Ulm University

Institute of Medical Systems Biology

Prof. Dr. Hans A. Kestler

SYSTEMS BIOLOGY OF TUMOUR EVOLUTION: ESTIMATING ORDERS FROM OMICS DATA

Dissertation submitted in partial fulfilment of the requirements for the degree of „Doctor rerum naturalium“ (Dr. rer. nat.) of the International Graduate School in Molecular Medicine Ulm

Lisa Monika Schäfer
Born in Bühl
Ulm, 2020

1. Current Dean of the Medical Faculty: Prof. Dr. Thomas Wirth

2. Chairman of the Graduate School: Prof. Dr. Bernd Knöll

3. Thesis Advisory Committee:

    - First supervisor: Prof. Dr. Hans A. Kestler

    - Second supervisor: Prof. Dr. Michael Kühl

    - Third supervisor: Prof. Dr. Paolo Frasconi

4. External reviewer: Prof. Dr. Dr. Steve Hoffmann

5. Further reviewer: Prof. Dr. Melanie Börries

6. Day doctorate awarded: 12.04.2021

This study was conducted with the support of Ontario Institute for Cancer Research through funding provided by the Government of Ontario. It provided one dataset.

Further acknowledgements have been removed due to privacy issues.

Results gained in my thesis have previously been published in the following publications:

- Lausser* L, **Schäfer* LM**, Schirra* LR, Szekely R, Schmid F, Kestler HA, Assessing phenotype order in molecular data. *Scientific Reports* 9, 1–10 (2019); https://doi.org/10.1038/s41598-019-48150-z
- Lausser* L, **Schäfer* LM**, Kühlwein S, Kestler AMR, Kestler HA, Detecting ordinal subcascades, *Neural Processing Letters* 52, 2583-2605 (2020); https://doi.org/10.1007/s11063-020-10362-0
- Lausser* L, **Schäfer* LM**, Kestler HA, Ordinal classifiers can fail on repetitive class structures, *Archives of Data Science, Series A* 4(1), 1-37 (2018); https://doi.org/10.5445/KSP/1000085951/25
- Lausser* L, **Schäfer* LM**, Kestler HA, Order constraints against the curse of dimensionality, *Studies in Classification, Data Analysis, and Knowledge Organization*, Springer (accepted)

*equal contribution

TABLE OF CONTENTS

List of figures VIII

List of tables IX

List of abbreviations, genes and symbols XI

1 Introduction 1

1.1 Levels of ordinality in biology 1

1.2 Phenotype level (macro scale of ordinality) 4

1.2.1 Ordinal classification 6

1.2.2 Ordinal classifier cascades 7

1.3 Signalling pathways (meso scale of ordinality) 8

1.4 Kinds of ordinality 10

1.5 Cancer 12

1.5.1 Cancer development 13

1.5.2 Pancreatic cancer 14

1.5.3 Colorectal cancer and hepatocellular carcinoma 16

1.6 Aims and overview 17

2 Methods and datasets 18

2.1 Notation 18

2.2 Classification 19

2.2.1 Learning a classifier 19

2.2.2 Terminology 20

2.2.3 Performance measures 20

2.2.4 Overfitting 21

2.2.5 Cross-validation 21

2.2.6 Binary classifier systems 22

2.2.7 Multi-class classification 25

2.2.8 Ordinal classification 26

2.2.9 Screening procedure 29

2.3 Reconstructing neighbourhood embeddings 30

2.3.1 Dependency measures 30

2.3.2 Self-organising maps 31
2.4 Network entropy 33
2.4.1 Protein interaction networks 34
2.4.2 False positive reduction strategies 35
2.5 Datasets and processing 40
2.5.1 High-throughput techniques and data processing 40
2.5.2 Gene enrichment analysis 41
2.5.3 Dataset characteristics 41

3 Ordinal classifier cascades & phenotype network reconstruction 47
3.1 Ordinal classifier cascade - advancements 47
3.1.1 Preselection strategy 48
3.1.2 Look-up structure - prediction map 50
3.1.3 Performance relations between cascades and their subcascades 51
3.1.4 R-package *ORION* 52
3.2 Ordinal classifier cascade - dependency on dimensionality 53
3.2.1 Experimental results 53
3.2.2 Theoretical results 54
3.3 Ordinal classifier cascade - characterisation 56
3.3.1 Simulating artificial data 56
3.3.2 Characterisation of the artificial data 59
3.3.3 Characteristics of the set of returned cascades 62
3.4 Phenotype network reconstruction 66
3.4.1 Reconstruction procedure 67
3.4.2 Impact of the bin size on the mutual information 68
3.4.3 Global and dataset specific thresholding 68
3.4.4 Reconstruction results 71
3.5 Discussion and conclusion 74

4 Characterisation of order types 77
4.1 Differentiation of types of ordinality 77
4.2 Experimental results 81
4.3 Discussion and conclusion 83

5 Results on real data screenings using the ordinal classifier cascade 84
5.1 Revealing orders in static data 84
5.1.1 Screening results 85
5.2 Cascades in developmental stages of model organisms 89
5.3 Cascades in datasets with (pre-)cancer stages 90
5.3.1 Screening settings 90
5.3.2 Parameter robustness of homogenous ordinal classifier cascade 92
5.3.3 Screening results 92
5.3.4 Analysis of possible candidate cascades 97
5.4 Discussion and conclusion 103

6 Network entropy highly depends on the protein-protein interaction network 110
6.1 Experimental setup 111
6.2 Effect of gene expression and network perturbations on network entropy 113
6.2.1 Effect of gene expression changes (knockouts) 113
6.2.2 Effect of correction methods for false positive edges 115
6.3 Network entropy courses of the (pre-)cancer stages 117
6.4 Discussion and conclusion 120

7 Conclusion 124

8 Summary 128

List of references 130

Appendix 140

LIST OF FIGURES

Figure 1 Biological dogma 2
Figure 2 Ordinality on various levels within an organism 3
Figure 3 Waddington landscape as a metaphor of the epigenetic landscape 4
Figure 4 Reflection of the order on the phenotype level in the gene expression level 5
Figure 5 Screening for ordinality by using the ordinal classifier cascade 7
Figure 6 Network entropy calculation scheme 9
Figure 7 The challenge of ordinality with increasing dimensionality 11
Figure 8 Different sample patterns might be considered as ordinal and distinguished by the characteristics of a base classifier 12
Figure 9 Evolution scheme of pancreatic adenocarcinoma 15
Figure 10 Three types of binary classifier systems 24
Figure 11 Skeleton of the ordinal classifier cascade 27
Figure 12 Example of decision region assignment using an ordinal classifier cascade 28
Figure 13 Pareto optimal solutions 29
Figure 14 Venn diagrams of protein and interaction number of three database resources individually and including the extracted strongly connected component of their union 35
Figure 15 Scheme of the applied techniques to correct for false positive edges in a protein-protein-interaction network 36
Figure 16 Overlap of pre-cancer stages within three analysed cancer types 43
Figure 17 Screening workflow of the ordinal classifier cascade 52
Figure 18 Classification performance dependency by increasing dimensionality for the ordinal classifier cascade and the *one-against-one* multi-class fusion architecture 54
Figure 19 Patterns of ordinal relations underlying the sampled artificial data 59
Figure 20 Pairwise prototype distances of linear datasets with 10 classes 60
Figure 21 Probability to sample a circle on the prototype level 61
Figure 22 Pairwise sample distances of linear datasets with 10 classes 62
Figure 23 Pairwise class separability with increasing levels of noise for all sampled patterns 63
Figure 24 Returned cascades for the linear pattern with 7 classes 64
Figure 25 Mean frequency of not expected cascades for all patterns 65
Figure 26 Scheme on the network reconstruction procedure 67

Figure 27 Dependency of the error rate on the bin size using mutual information as dependency measure 68
Figure 28 Average error rate per pattern group and method dependent on the global threshold 70
Figure 29 Error rates of reconstructing patterns using a global threshold of 0.3 72
Figure 30 Construction scheme of a component graph using a 4-class example 79
Figure 31 Example scheme of the four ordinality criteria 80
Figure 32 Prototype locations of the diverse sampled artificial ordinal patterns 81
Figure 33 Performance of the returned cascades in diverse artificial presumable ordinal patterns 82
Figure 34 Cascade results on the hypothalamic-pituitary-adrenal-axis dataset (d_1) 87
Figure 35 Cascade results on the leukaemia dataset (d_2) 88
Figure 36 Overview of the analysed screening settings 91
Figure 37 Base classifier performance on the pre-cancer datasets for linear L1-support-vector-machine and a homogenous cost parameter 97
Figure 38 Classification of external classes for the returned pre-cancer cascades 98
Figure 39 Histogram on the selected features for each finally selected cascade of the pre-cancer datasets 99
Figure 40 Overlap of the selected features of the base classifiers that make up the finally selected cascades of the pre-cancer datasets 100
Figure 41 Scheme of the experimental setup for network entropy evaluation 112
Figure 42 Network entropies of simulated phenotypes 113
Figure 43 Impact of three perturbation types on the network entropies 114
Figure 44 Correction technique influence on true network characteristics 116
Figure 45 Performance of false positive correction techniques 117
Figure 46 Entropy results on four pre-cancer datasets 118
Figure 47 Significant pairs of classes using a Mann-Whitney-U test with Holm correction 119

LIST OF TABLES

Table 1	Characteristics of the real datasets	42
Table 2	Look-up table for the preselection of possible orderings	50
Table 3	Prediction map memoizing the predictions of all binary classifiers	51
Table 4	Rate of found not expected cascades within the set of patterns without relations	66
Table 5	Characteristics of the candidate set comprising the longest cascades of static datasets	85
Table 6	Set of returned cascades for datasets of developmental stages of model organisms	89
Table 7	Robustness of returned cascades screening the pre-cancer datasets against the cost parameter using a linear L1-support-vector-machine homogenous ordinal classifier cascade	93
Table 8	Screening results on the pre-cancer datasets selecting the Pareto optimal solutions	94
Table 9	Results on Kyoto Encyclopedia of Genes and Genomes (KEGG) gene enrichment analysis of the colon carcinoma dataset (d_7)	101
Table 10	Robustly selected features in the pancreatic adenocarcinoma cascades (d_8) that represent genes	102

LIST OF ABBREVIATIONS, GENES AND SYMBOLS

ABBREVIATIONS

ALL	acute lymphocytic leukaemia
AML	acute myeloid leukaemia
aOcc	accuracy tuned ordinal classifier cascade
APC	adenomatous polyposis coli
BioGRID	Biological General Repository for Interaction Datasets
cDNA	complementary deoxyribonucleic acid
C. elegans	*Caenorhabditis elegans*
CLL	chronic lymphocytic leukaemia
CML	chronic myeloid leukaemia
COMPASS	Comprehensive Molecular Characterisation of Advanced Pancreatic Ductal Adenocarcinomas for Better Treatment Selection
CRAN	Comprehensive R Archive Network
CV	cross-validation
CD	Czekanowski-Dice
DAG	directed acyclic graph
DNA	deoxyribonucleic acid
FDR	false discovery rate
FN	false negative
FNR	false negative rate
FP	false positive
FPR	false positive rate
GEO	Gene Expression Omnibus
GNER	global network entropy rate
GO	gene ontology
HBV	hepatitis B virus
HCC	hepatocellular carcinoma
HCV	hepatitis C virus
hOcc	homogenous ordinal classifier cascade
HPA	hypothalamic-pituitary-adrenal
invlog	inverse log-weighted
KEGG	Kyoto Encyclopedia of Genes and Genomes
k**NN**	k-nearest neighbour
LNE	local network entropy
LL	lymphocytic leukaemia
LOOCV	leave-one-out-cross-validation
MDS	myelodysplastic syndromes
MI	mutual information
ML	myeloid leukaemia
MLL	mixed-lineage leukaemia
mRNA	messenger ribonucleic acid
NN	nearest neighbour
OaO	*one-against-one*
Occ	ordinal classifier cascade
PanIN	pancreatic intraepithelial neoplasm
PC	Pathway Commons
PCR	polymerase chain reaction
PDAC	pancreatic adenocarcinoma
PPI	protein-protein interaction
rF	random forest
rma	robust multi-array average
RNA	ribonucleic acid

RPKM	reads per kilo base per million
scc	strongly connected component
SCCB	squamous cell carcinoma of urinary bladder
sOcc	sensitivity tuned ordinal classifier cascade
SOM	self-organising map
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
SVM	support vector machine
TG	tumour grade
WHO	World Health Organisation

GENES

BRCA2	BRCA2, DNA repair associated
CACNG4	calcium voltage-gated channel auxiliary subunit gamma 4
CDKN2A	cyclin dependent kinase inhibitor 2A
CTNNB1	catenin beta 1
DAXX	death domain associated protein
DCK	deoxycytidine kinase
DNASE1	deoxyribonuclease 1
DGUOK	deoxyguanosine kinase
ERBB2	erb-b2 receptor tyrosine kinase 2
HIF	hypoxia-inducible factor
HOXA4	homeobox A4
IGF2/H19	insulin like growth factor 2/H19 imprinted maternally expressed transcript
KIAA1805/ ZNF512	zinc finger protein 512
KRAS	kirsten rat sarcoma viral oncogene homolog
KRT1	keratin 1
KRTAP	keratin associated proteins
NR3C1	nuclear receptor subfamily 3 group C member 1
NXPE4	neurexophilin and PC-esterase domain family member 4
MAPK	mitogen-activated protein kinases
MDK	midkine
PGC	progastricsin
PIGZ	phosphatidylinositol glycan anchor biosynthesis class Z
PIP5K1B	phosphatidylinositol-4-Phosphate 5-Kinase Type 1 Beta
PNLIPRP2	pancreatic lipase related protein 2 (gene/pseudogene)

RXRG/ NR2B2	retinoid X receptor gamma
SMAD4	SMAD family member 4
SMARCC2	SWI/SNF related, matrix associated, actin dependent regulator of chromatin subfamily c member 2
SPN/ CD43	sialophorin
TP53	tumor protein p53
XIST	X inactive specific transcript

SYMBOLS

$\prec$	ordinal semantic relation
$y_i \prec y_j$	ordinal semantic relation between labels y_i, y_j
$y_{(i)} \prec y_{(j)}$	ordinal semantic relation between position i, j
$\circ$	Hadamard-product
a	number of units in map
$acc_c(y)$	accuracy of y using c
A	adjacency matrix
A_{PPI}	adjacency matrix of PPI
α, β	parameter of Γ-distribution
b	hyperplane shift
c	binary classifier
$c_{(i)}$	binary classifier at position i
c_{ij}	binary classifier trained on labels y_i, y_j
cca	closest common ancestor
cov	covariance
cp	component
C	cost parameter (SVM)
$\mathcal{C}$	set of candidate cascades
CC	correlation coefficient matrix
CP	set of components
d	distance
de	degree distribution
deg	degree
deg^{in}, deg^{out}	in-degree, out-degree
$\mathcal{D}$	decision region
D	set of distances
δ	set of adherence points/ boundary
e	Euler number
$\mathbf{e_{ij}}$	edge between i, j
err	error rate
E	expression product
$\mathcal{E}$	ensemble of base classifiers
E	set of edges
EP	number of expected and detected
ϵ	learning rate
f	function
f	fold change vector
f_k	kernel function
f_o	function for number of possible cascades
f_{rel}	relative frequency distribution
FC_{ij}	sensitivity of first class y_i using c_{ij}
g	gene
G_c	component graph of classifier c
$\mathcal{G}$	graph
G	gene set

γ	parameter in degree distribution
Γ	gamma distribution
h	multi-class classifier scheme
H	entropy
ic	information content
i, j, l	running indices
k	number of labels
κ	weighting parameter
len_g	length of gene g
λ	Lagrange multipliers
m	sample size
$\tilde{m}$	number of bins for discretisation
$\hat{m}$	number of non-zero bins
MI	mutual information
μ	mean
n	feature number/ dimensionality
N, N°	number of
$\mathcal{N}$	normal distribution
$\mathbf{N}(g)$	set of first-level neighbours of g
$\mathbf{N}^{anc}(g)$	set of first-level ancestors of g
$\mathbf{N}^{desc}(g)$	set of first-level descendents of g
NE	number of edges that do not exist
$NEDR$	not expected discovery rate
NEP	number of not expected but detected
$NEPR$	not expected positive rate
ν	parameter cumulative normal distribution
o	number of pairwise relations
osv	overall semantic value
$\mathcal{O}$	set of classes
$\mathbf{O}$	set of ordinal relations
p	one feature in prototype vector
$\tilde{p}$	curve/ function
p_c	prediction rate using c
p_{ij}	transition probability
$\mathbf{p}$	prototype vector
$prob$	probability
$\widetilde{prob}$	estimated/empirical probability distribution
$pur_c(y)$	purity of y using c
P	transition matrix
$\mathcal{P}$	partition
$\mathbf{P}$	set of prototypes
PE	number of edges that exist
PPI	directed graph of protein-interaction network
Φ	cumulative normal distribution
π	stationary distribution
q	polynomial degree
Q	optimisation function (SVM)
r	regularisation parameter (SVM)
$\mathbf{r}$	direction regulation vector
rk	rank function
R	reliability/similarity score
R^{sem}	semantic reliability/similarity score
R^{top}	topological reliability/ similarity score
ρ	correlation coefficient
$sens_c(y)$	sensitivity of y using c
$smin$	minimal class-wise sensitivity
sv	semantic value

S	local network/ signalling entropy	$\mathcal{U}$	uniform distribution
$SC_{ij}(l)$	prediction rate of second class y_j using c_{ij} and a sample of class y_l	$\mathbf{V}$	set of labelled validation data
		$\mathbf{w}$	multi-dimensional angle
		ω	Euclidian topology
SR	global network/ signalling entropy rate	x_i	one feature
		$\tilde{x}$	number of samples per bin
σ	standard deviation		
$\sigma_{mult}, \sigma_{add}$	multiplicative/ additive noise factor	$\mathbf{x}$	sample vector
		$\mathcal{X}$	input space
Σ	covariance matrix	$\mathbf{X}$	sample set
t	threshold	$\mathbf{X}_y$	labelled sample set
tm	term in taxonomy	$\mathbf{X}_y$	sample set with samples of label y
$\mathbf{T}$	set of labelled training data	ξ	slack variable
$\mathbf{TM}$	set of terms	χ^2	χ^2 distribution
TX_{map}	topological map for feature space	$y_{(i)}$	label at position i within the cascade
TY_{map}	topological map for label space	y_i	label y_i
		$\mathcal{Y}$	label set
τ	distance threshold of update in SOM	$\mathbf{z}$	sampled centroid
		ζ	semantic contribution factor
u	iteration counter		

1

INTRODUCTION

In this introduction I will briefly give an overview of biological processes and relations between biological components that might be considered as ordinal (Section 1.1). I will first introduce relations on the phenotype (macro scale) level and a method, which might be suitable to study those: ordinal classification and more specifically ordinal classifier cascades (Section 1.2). Afterwards, relations on a meso scale level (signalling pathways) as well as network entropy a method that uses ordinality on this level to study orders on the phenotype level is shortly explained (Section 1.3). Furthermore, the question of the meaning of ordinality in high-dimensional space is tackled (Section 1.4). And finally, the current knowledge and terminology regarding cancer research in general and pancreatic and colorectal cancer as well as hepatocellular carcinoma in particular are explained (Section 1.5).

1.1 Levels of ordinality in biology

Well-orchestrated processes determine the functionality on each level of the human body. The interplay between processes within a cell, between cells and between organs make the human organism function as a whole. Within a cell the amount, composition and interplay of genes, proteins and metabolites determines the cell's functionality [96]. Therefore, analysing the genome, the transcriptome and the proteome and more recently the epigenome, the metabolome and the microbiome together, but also on their owns lead to a better understanding of how cells work [77, 49, 209, 97] and why diseases occur [124, 40, 171]. The transcriptome hereby is mainly thought to serve as connective between the genome and the proteome. The information stored in the genome is first transcribed to messenger ribonucleic acid (mRNA) that makes up the transcriptome. Afterwards, it is translated to proteins that perform many tasks within a cell [52] (Figure 1). The correlation between the transcriptome and the proteome is still under debate [224, 132, 218]. It might be that the poor correlation, which is often denoted [3] is rather due to systematic measurement biases or high stochastic variation [224, 127]. Studies exist, which show that the variation on the transcript level can be considered as corresponding to the

protein concentration differences [126, 189]. Independent of whether there is a high correspondence between the transcript level and the protein level, the gene expression level can be used to get insights into the cell's state and into intra-cellular processes in healthy and diseased states [172, 250, 43, 136, 8].

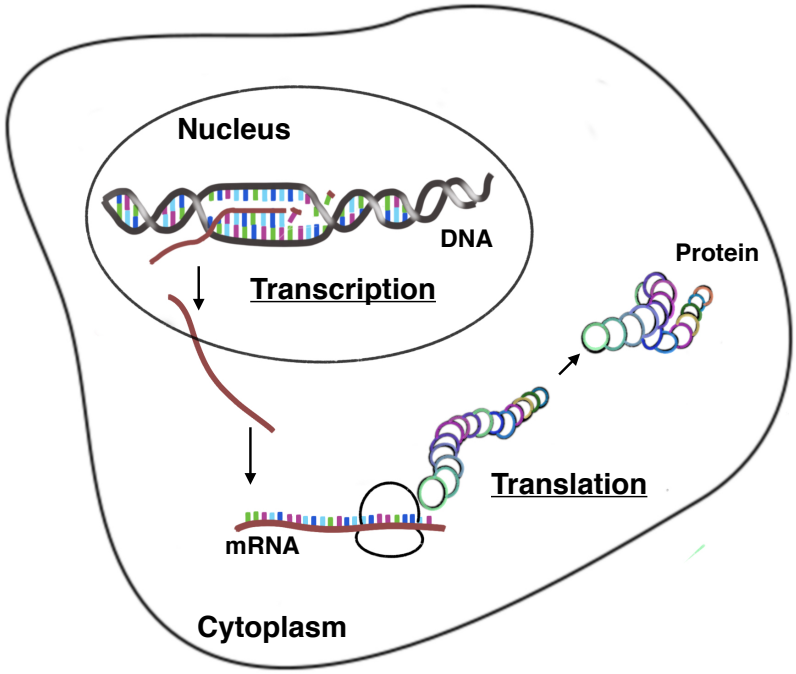


Figure 1: Biological dogma. The genetic information stored in the nucleus using deoxyribonucleic acid (DNA) is transcribed to the messenger ribonucleic acid (mRNA). The mRNA is translocated to the cytoplasm und translated to an amino acid sequence that makes up protein.

Mathematically spoken one deals with dynamical processes and the measured cell states or the whole tissue phenotypes can be considered as stable states of the dynamical interplay [107, 235]. Usually, fluctuations or less severe changes within the processes do not lead to changes in the macro scale, the observed characteristics. The functionality of the cell is not disturbed in the long term, as the system itself is robust and converges to the original state [231, 110, 54]. However, if the disturbances are too large, like changes in the external environment [167], multiple mutations [135, 192] or chromotripsis, which is the simultaneous rearrangement of several parts of the genome [183], the system travers to another stable state. These transitions can be demanded, if performed in a controlled manner, like in embryonic development or during cell differentiation [14, 48]. However, in the same way they often lead to diseases like among others cancer [12, 36, 245]. Hereby, not each abnormal phenotype might be a risk for the whole organism, as for example there are neoplasms showing an abnormal phenotype but are considered being

non harmful [22]. The process of a normal healthy phenotype towards an abnormal harmful phenotype might occur by a sequence of stable states [12] or by a one step transition (chromotripsis) [183].
Overall, if studying the transition from normal to harmful phenotypes, two levels of ordinality play an important role. There is on the one hand the meso scale level, like the order of proteins within a signalling cascade and on the other hand the order on a macro scale level, the order between observable stable states (Figure 2).

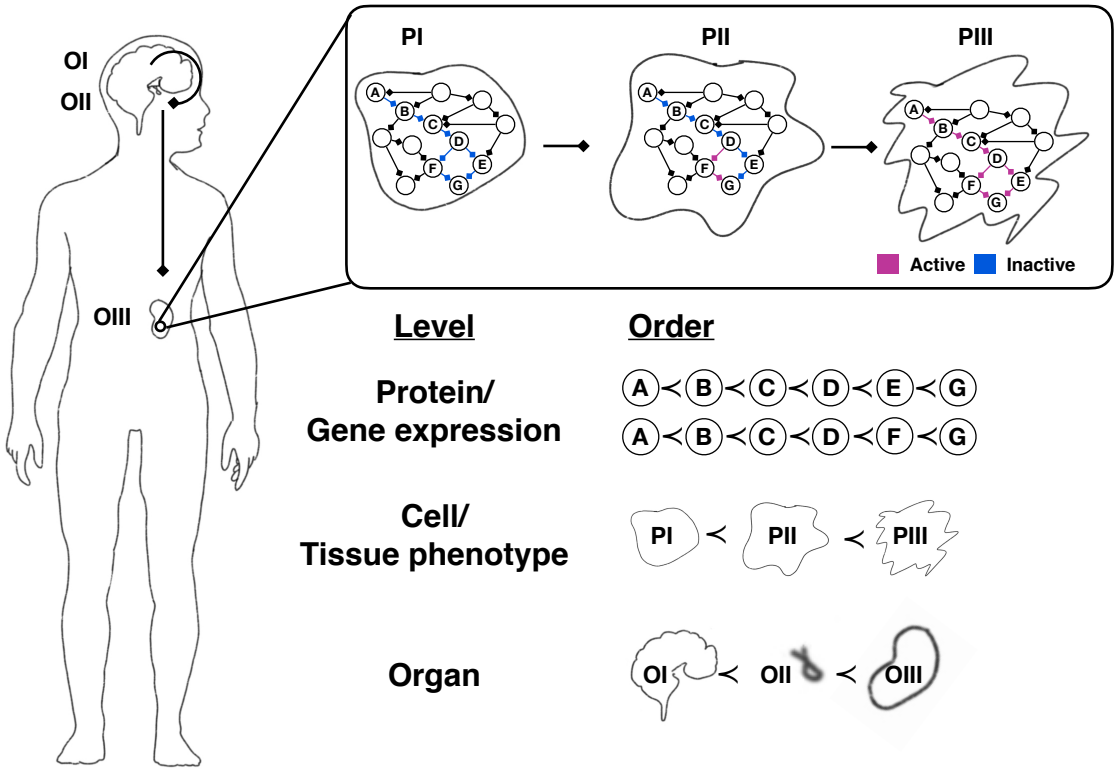


Figure 2: Ordinality on various levels within an organism. An organism is a complex system functioning by an interplay of several and diverse processes. Functioning units can be grouped. This grouping can be done on different scales and on each scale ordinal relations might be found. On a broad scale one can deal with organs. Zooming in, a tissue and cell scale level might be considered and these functional units are themselves made up of relations between genes and proteins. One example on the organ level (OI, OII, OIII) might be the hypothalamic-pituitary-adrenal (HPA)-axis describing a relation between the hypothalamus, pituitary and the adrenal glands. On the cell level (PI, PII, PIII) a relation between developmental phenotypes or pre-cancer stages can be seen. The smallest scale depicted here is the protein level (A, B, C, ...). Here many relations between proteins can be found. Furthermore, these levels build upon each other, as the kind of relation on the protein level (active/ inactive) might determine the relation on the cell level. The cell or tissue level itself might play a role on the organ level.

1.2 Phenotype level (macro scale of ordinality)

Waddington [221] proposed a metaphor to describe the relationship between the gene activity and the resulting phenotype during development. This epigenetic landscape might be used to illustrate the transitions on the phenotype level. The landscape is defined by the activity of genes and comprises all possible gene activity combinations. He compared cell differentiation with the process of a ball rolling down a hill and ending up in one of several valleys. These valleys correspond to the final cell states. The different valleys are separated by ridges and hence the ball cannot move to another valley without an external force (Figure 3). This metaphor goes along with the descriptions of cell fate decisions in systems biology. Independent of the model type used, Boolean models or models based on ordinal differential equations, in simulations it is searched for steady states or attractors, referring to biological phenotypes, trajectories describing potential transitions and bifurcation points at which cell fate decisions are made [155]. Hereby the actual potential measure is not explicitly included.

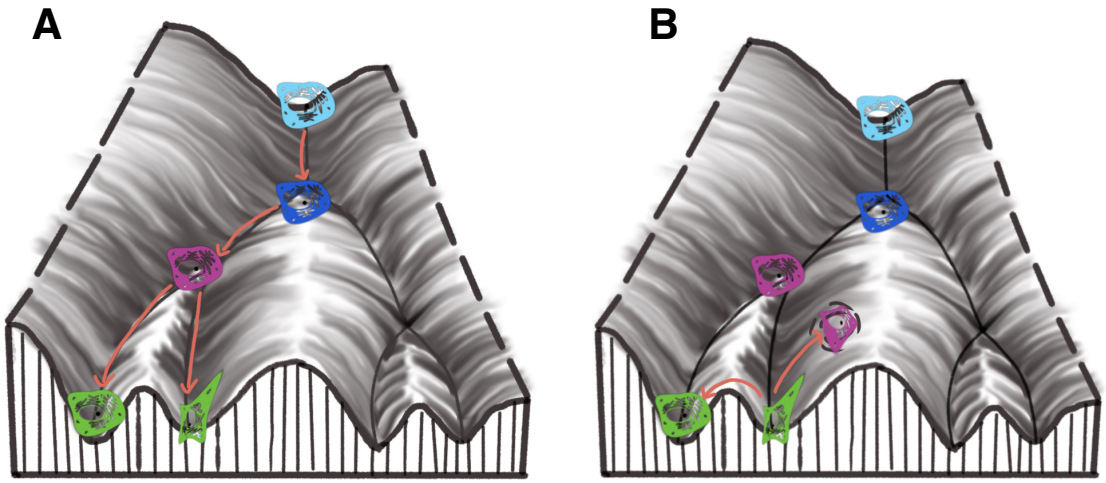


Figure 3: Waddington landscape as a metaphor of the epigenetic landscape. This metaphor depicts the idea that cell development can be compared to the movement of a ball on a landscape. A: Starting as stem cell (light blue) on the top, the cell moves downward while differentiation and there exist bifurcation points (dark blue, violett), which determine the cell fate decision. The differentiated cells are located in valleys, each valley representing one cell type (green). B: If there is a transition from one differentiated cell type towards another or if new abnormal cell types occur the cells have to pass a barrier.

As the space spanned by this landscape contains all possible cell states, not only the cell controlled process of differentiation, but also transitions occurring due to external forces or due to large stochastic fluctuations, like the transition of normal towards tumour cells, can be thought of as a movement on this landscape. Small fluctuations might cause the cell return to its previous valley, a characteristic referred to as robustness [231, 110, 54]. Otherwise the cell will move to another valley potentially leading to a change in phenotype.

Under the assumption that it is more likely that a cell travers to a neighboured stable state, there is an order on the phenotypes and this order should also be reflected in molecular characteristics (Figure 4).

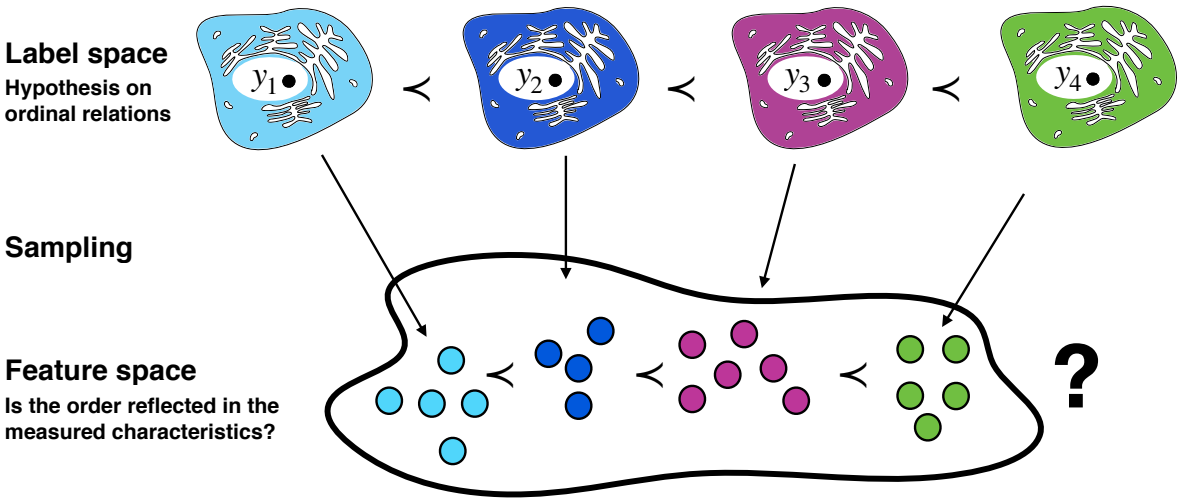


Figure 4: Reflection of the order on the phenotype level in the gene expression level. The order defined on the phenotype level (label space) is assumed to be reflected in the gene expression (feature space). Depicted is an example with four phenotypes (y_1, y_2, y_3, y_4). These phenotypes might be characterised by their appearance, which also lead to a hypothesis about an ordinal relation between them. Measuring samples determines diverse characteristics spanning the feature space. It is often assumed but not shown that the order hypothesised on the labels is also given in the feature space. This might therefore be questionable. See [123] licensed under CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

Especially for diagnostic reasons it is essential to define well-distinguishable states, as each state might lead to different treatment options [11, 22, 206, 257]. Additionally, knowing not only the direct neighbourhood of these stages, but also the sequence of orders, depicting transition processes, might guide treatment decisions. As one focus is on the distinction between the stages ordinal classification, combining the functionality of finding well-distinguishable classes, interpreted as observable phenotypes, with the constraint of an order on these classes can be used to

study the order of phenotypes on the phenotype level (label space) and its reflection in the molecular characteristics (feature space).

1.2.1 Ordinal classification

Learning a classifier aims at finding a general rule that best separates samples labelled differently. Subsequently, these rules can be applied on new, unseen samples. Hereby, binary classification deals with two labels and multi-class classifier systems are used for tasks with more than two classes [84].
Ordinal classification is a constrained multi-class classification problem that takes into account the order information given in the label space $y_1 \prec y_2 \prec \ldots \prec y_{|\mathcal{Y}|}$. The symbol $\prec$ is used as an order relation on a semantic/qualitative and not a quantitative level and $\mathcal{Y}$ denotes the label space. The additional knowledge may guide the classifier and this is assumed to lead to a better performance. The value of the order information might be seen as the performance increase as the ordinal structure should be reflected by the class distributions [95].

Similar to a general multi-class classification problem, ordinal classification tasks can be tackled by binary decomposition. Binary decomposition splits the multi-class classification problem into binary classification tasks that are combined by a fusion architecture, like *one-against-one* (OaO), *one-against-all* [27] or ordinal classifier cascades [119]. Further approaches might be a naive approach using standard machine learning techniques by for example ignoring the order and using nominal classification, or by transferring the categories into real values and applying regression. These methods have to deal with either the fact to not include given knowledge or problems like the unknown distance between labels or the decision of which real value is assigned to which label [80].

Using a binary decomposition scheme, an often employed combining strategy is cascading. It was introduced by Viola and Jones, who arranged the binary classifiers based on their strengths [217]. Cascading refers to the strategy that binary classifiers are concatenated and all information of the previous classifiers are given to the next classifier. An example for a cascading fusion scheme are classifier chains, a multi-label classifier system. The binary classifiers are arranged in a chain and the result of each classifier is given as features to the next classifier accounting for correlations between the classes. To overcome the dependency on the chain order, ensembles of classifier chains are used [181]. Cascading is also employed by arranging binary classifiers as a directed acyclic graph [173] or by building trees of binary distinctions of subgroup splits, as in nested dichotomies [69]. Further ordinal classifier systems are the method introduced by Frank and Hall [68] that exploits the

order information by an encoding in binary classifiers that can predict probabilities or the data replication method searching for parallel hyperplanes by requesting non-intersection boundaries [38]. Constraining the order assumption to an increase or decrease in all features the term monotone classification is used [65].

1.2.2 Ordinal classifier cascades

The ordinal classifier cascade (Occ) belongs to the multi-class classification systems using binary decomposition and a cascading scheme as fusion architecture. In a screening experiment on small artificial datasets it was shown that using a pairwise training scheme the Occ shows a diminished performance if the assumed order is not reflected in the investigated feature space [119]. The Occ is built as a sequence of binary base classifiers, which are evaluated sequentially. If one base classifier predicts its first class the label is determined and the cascade stops. Otherwise the next base classifier is evaluated. A detailed description together with all further developments conducted within the scope of this thesis is given in Sections 2.2.8.1 and 3.1.
The diminished performance in addition to the pairwise training scheme and a preselection procedure, of which the functionality was shown within this thesis (Section 3.1) allow for using the ordinal classifier cascade in large scale screening experiments (Figure 5) (Section 5.1) [123, 122].

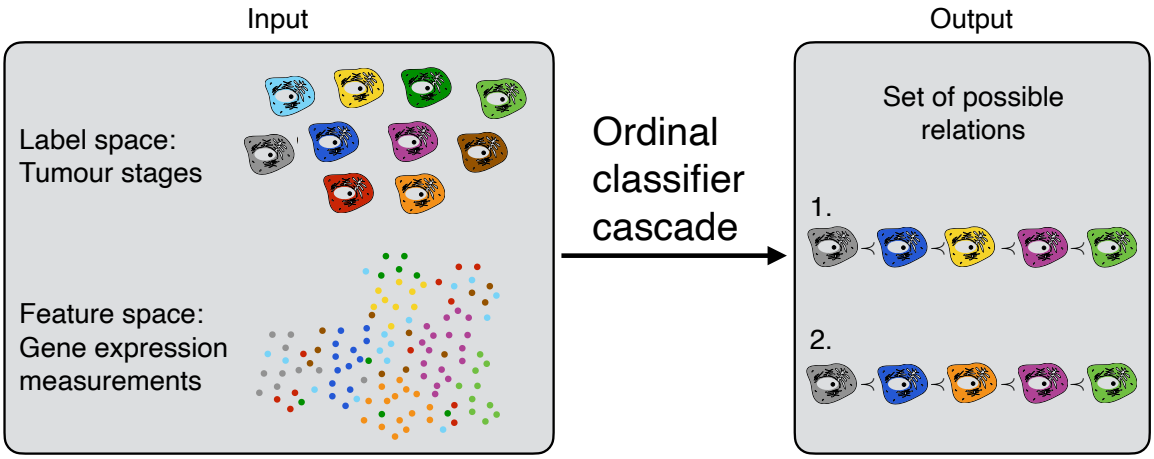


Figure 5: Screening for ordinality by using the ordinal classifier cascade (Occ). Given labelled data as input, for example tumour stages and their corresponding gene expression measurements, the Occ screens through this information and outputs a candidate set of possible relations that could be revealed by screening through the feature space. Additionally, the candidates might be ranked based on their performances.

By this the Occ largely differs from the previous mentioned ordinal classification techniques [181, 173, 69, 68, 38, 65]. It cannot only be exploited for improving classification performance by the use of ordinal information but also for searching for or confirm potential orders within a given feature representation. If Occ is used to screen for reflected orders in the feature space it is used in order to analyse data intrinsic structures. The reflected orders might either be used to improve classification performance [95] or to gain further insights into the research context [123, 122].

1.3 Signalling pathways (meso scale of ordinality)

As mentioned in the previous section the cell or tissue phenotype is determined by its molecular characteristics described as building the epigenetic landscape by Waddington [221]. Hence, the transition between phenotypes goes along with changes in the intracellular characteristics and dynamics. The intracellular regulations rely on the gene regulatory or protein-protein interaction network, a complex network characterised by a variety of signalling pathways. A reference for signalling pathways can be found in Kyoto Encyclopedia of Genes and Genomes (KEGG) [105]. Signalling information is committed along paths in an ordered manner. As such, the relation between gene expression changes in ordered signalling pathways and the corresponding phenotypes is of interest [178]. A combined measure of interaction information, determined by the pathways, together with changes in the expression might give new insights into the cells' phenotypes and their relations.

One measure that combines these types of information is network entropy, or network entropy rate [56]. The interaction information is considered as a priori knowledge and extracted from databases. The gene expression is derived from experimental measurements and can be used to estimate interaction strengths for the interactions extracted from databases. Based on the resulting weighted interaction network the network entropy is calculated (Figure 6) [18].
The local entropy claims for each node how uncertain the next step is having arrived at that node. The equilibrium global network entropy describes the distribution of information flow in the whole network and might be considered as a measure for the complexity of the system [18]. A detailed description on the network entropy calculation can be found in Section 2.4.

Network entropy rate measures the information flow within a network. Comparing two phenotype stages and assuming the same static intracellular network topology this measure indicates whether the network flow is distributed differently. For example, in the case of gain-of-function or loss-of-function mutations the flow is in-

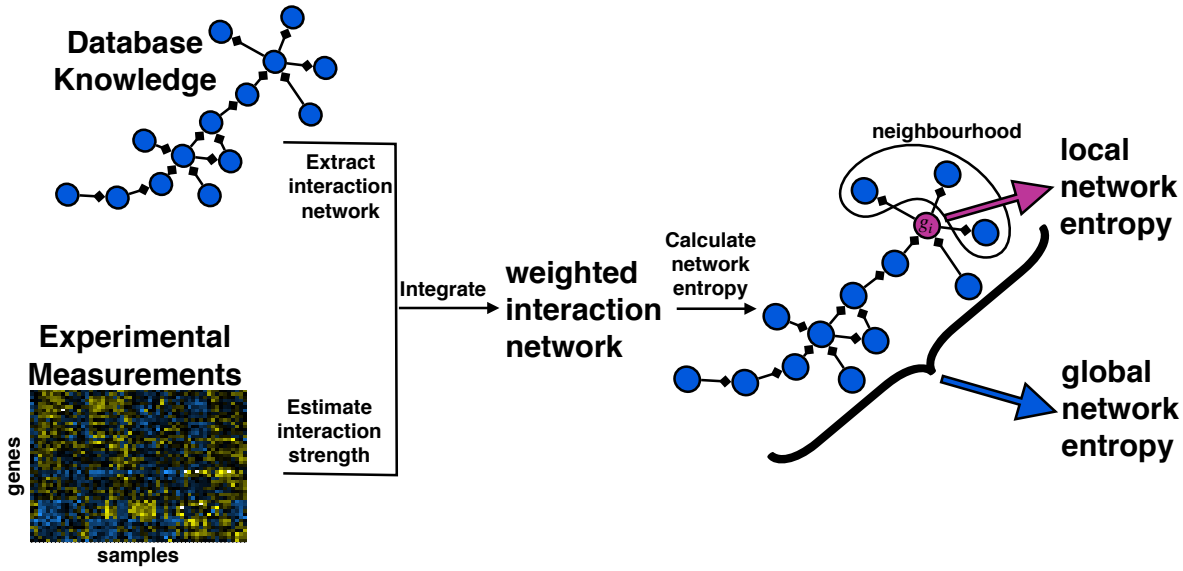


Figure 6: Network entropy calculation scheme. The network entropy calculation requires the fusion of the global information given by a protein-protein- or gene-interaction network stored and collected in diverse databases and the information of specific experimental measurements, like gene expression data. Based on the mass action principle interaction strengths are calculated, used as basis to measure interaction probabilities and then combined with the extracted interaction network. The resulting matrix is used to calculate the local and global network entropy.

creased or decreased, respectively. Hence, it can serve as a potential measure on the Waddington landscape [221]. Demetrius and Manke [56] further showed that this measure represents the robustness of a system, defined as the rate of convergence. If the rate of convergence is high the system converges faster towards the unperturbed state and is hence claimed as more robust than systems that need longer to converge. In this sense robustness can be understood as the property of a system to withstand perturbations. Furthermore, it was shown that robustness is positivly correlated with entropy [57] leading to the conclusion that more robust networks have a higher global network entropy rate.

The theory of network entropy has already been applied to tackle various biological and medical questions. In 2010 Teschendorff and Severini [205] analysed four different breast cancer data sets by combining them with a protein-interaction network derived from the Human Protein Reference Database [109] and found a positive shift in the local entropy distribution for metastatic cases, which is interpreted as an increase in local randomness.

A similar study was performed by West et al. [230]. They used the network entropy to compare normal and tumour tissue for different cancers, like bladder, lung, gastric, pancreas, cervix and liver, and also showed a shift towards higher entropies. This was confirmed by Cheng et al. [46] in 2016 using ribonucleic acid (RNA)-sequencing data from eight different cancer types. They additionally investigated local entropy distributions of four consecutive tumour states. The entropy distributions between the four analysed tumour states differed, however not in a consistent way between all tumours. Whereas there is a decrease observed in the stage IV tissue for breast invasive carcinoma, colon adenocarcinoma and lung adenocarcinoma, the local network entropy increases for head and neck squamous cell carcinoma. In contrast to the previous study that concentrated on the local network entropy distribution, Banerji et al. [18] showed that the global network entropy rate increases between normal and cancer tissue, whereas it decreases while differentiation. This led to the previously mentioned hypothesis that this measure can serve as a quantitative potential measure on the Waddington landscape of cell differentiation [221]. Furthermore, network entropy could be related to clinical outcome in breast and lung cancer [19]. Looking at cancer predisease states, Sato et al. [185] figured out that there is a monotone increase in the global network entropy rate with the order of preneoplastic lesions of cervical carcinoma.

1.4 Kinds of ordinality

Being ordinal refers to being able to claim relative positions. Ordinality is determined by a set of relations defined on the objects that fulfil the requirements of transitivity, reflexivity and antisymmetry [190]. The interpretation and the position of an object usually depends on the characteristics of the objects that are ordered. If the characteristics are measured numerically and only one characteristic is considered, an order is usually defined by either increasing or decreasing values (Figure 7A) and the interpretation is given by the attribute. Using different characteristics of the same objects the relative position of these objects might differ for each corresponding attribute (Figure 7B). Furthermore, it can be distinguished between total orders, an order on all classes, or orders on subsets of classes (Figure 8A, B).

Using the examples shown in Figure 7 it can be seen that already in two dimensions an order considering all characteristics is neither clearly defined nor easy to interpret. Increasing the dimensionality many possible orderings and in the same way many geometrical arrangements can occur that might still be considered as ordinal (Section 4) (Figure 7C).

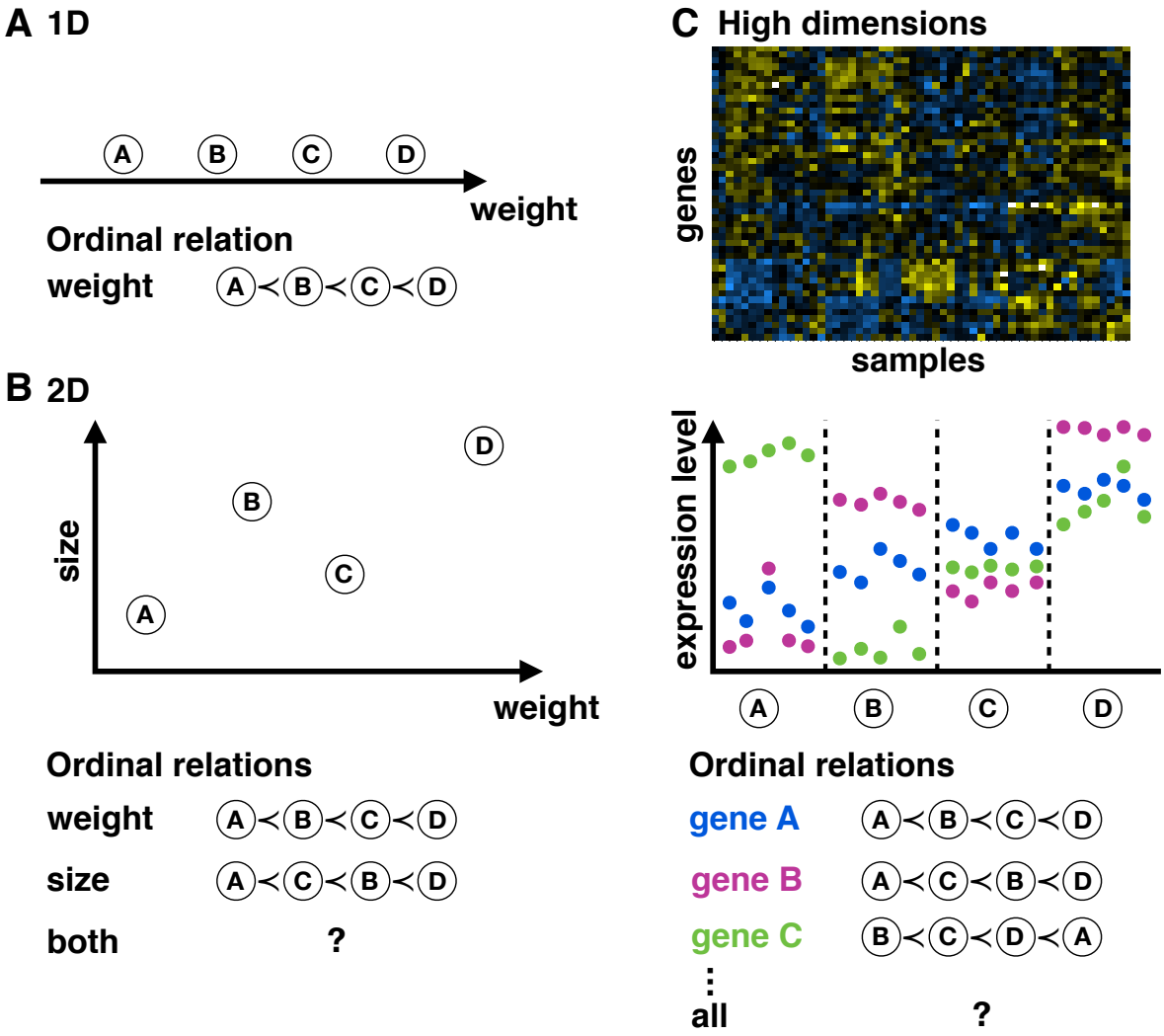


Figure 7: The challenge of ordinality with increasing dimensionality. A: An ordinal relation in one dimension is usually characterised by the one attribute itself. B: Increasing the dimensionality there may be different orders depending on the analysed attribute. An ordinal relation considering both attributes might be not clearly defined. C: The possibilities of ordinal relations increase with high dimensions, like for example gene expression data. There might be orders on each single attribute or orders on subgroups. Also diverse geometric arrangements might be considered as ordinal. The meaning and interpretation of an ordinal relation considering all attributes is not distinct.

To reveal orders from high-dimensional space the ordinal classifier cascade can be used, whereas the type of base classifier might lead to different decision regions and possibly orders of different interpretation (Figure 8C). Another kind of order might be found by the working definition for ordinal classification of Bellmann and

Schwenker [25]. They base their definition on the pairwise accuracy of classes as measure of separability and consider an order as given in the data representation, if only this one order leads to an ordered sorting on the pairwise accuracy matrix. Furthermore, orders might be found exploiting the data depth concept by Tukey [212]. It describes the centrality of a sample in a sample cloud and one might refer to this centrality measure in order to conclude an order on the data under investigation.

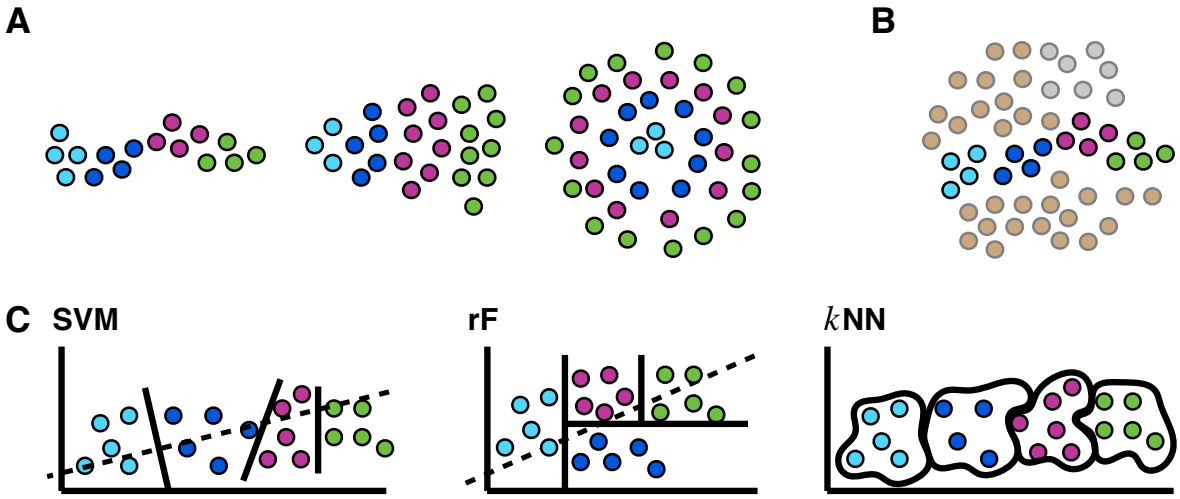


Figure 8: Different sample patterns might be considered as ordinal and distinguished by the characteristics of a base classifier. Each circle corresponds to one sample and the colour determines its class. A: In two dimensions ordinality in the feature space might be determined by different patterns. B: There might be an order within a subset of classes. C: The characteristics of the decision regions drawn by different base classifiers (support vector machine (SVM), random forest (rF), k-nearest neighbour (kNN)) might be used to distinguish between kinds of ordinal patterns.

The revealed order might be interpreted as a time line, a spatial order or even as dependencies. The interpretation of the found order is based on the subject under investigation and often needs additional background knowledge. This is in contrast to causal inference for which many measurements of time series data are required [168], and which is therefore not suitable for finding orders in phenotypes taken at the same biopsy.

1.5 Cancer

Cancer is characterised by dynamic changes in the genome leading to abnormal cell and tissue characteristics. Main characteristics, known as the seven hallmarks of

cancer are tissue invasion and metastasis, limitless replicative potential, sustained angiogenesis, evading apoptosis, self-sufficiency in growth signals and insensitivity to anti-growth signals [81]. Estimations by the World Health Organisation (WHO) show that in more than a half of 172 investigated countries it ranks first or second in the causes of death before an age of 70 and it is expected to become the global leading cause of death of the 21st century [31].
In general it holds that the earlier cancer is detected the better it can be treated and the better the prognosis [85]. This makes it essential to investigate and understand the development of cancer in a detailed but also conceptual manner, and to be able to estimate the risk or progression status of pre-cancer lesions.

1.5.1 Cancer development

A suitable and precise categorisation is needed for decision making in diagnosis. Based on histological observations and sample analysis the patient's stage is categorised, which serves as the basis for prognosis and potential treatments. The defined categories are usually associated with the severity of the disease and often serve as basis for the progression of the disease. In general there are potential pre-cancer stages and cancer stages. The pre-cancer stages are often based on histological observations and ordered according to the amount of differences from normal tissue. They are categorised based on their epithelial architecture and the characteristics of the nucleus [55]. The following terms are used to describe the histological pre-cancer observations and stages. They are ordered based on their presumed grade of alteration:

- **neoplasia**: cells that have an appearance and proliferation pattern different from normal [149].
- **dysplasia**: abnormal growth and differentiation [55].
- **hyperplasia**: the increase in size of a tissue or organ caused by for example an increase in the cell number (in contrast to an increase in size due to larger cells (hypertrophy)) and the opposite of atrophy (the effect that tissue decreases) [149].
- **metaplasia**: differentiated cells of a specific type are replaced by differentiated cells of another type, the tissue type is transformed [149].
- **adenoma**: in epithelial tissues, premalignant non invasive growth (benign) [149].
- **carcinoma in situ**: abnormal cells but non invasive, reside at their location, no metastatic potential [70].
- **adenocarcinoma**: tumour developed from secretory epithelial cells [149].

1.5.2 Pancreatic cancer

Pancreatic cancer is one of the cancer types that goes along with a bad prognosis and a short survival time. Globally, it is the seventh leading cause of cancer deaths [31]. In the United States it makes just 3-4 % of the newly reported cases of cancer, but it makes up 7 % of the cancer deaths. Only around 8 % of the diagnosed people survive the first 5-years after the diagnosis. In 1975 the 5-year survival rate was 3 %, which leads to a 5 % improvement within around 35 years and shows that it is still a challenge to treat that cancer [91]. There are different kinds of pancreatic cancer, on the one hand the pancreatic adenocarcinoma (PDAC), accounting for about 90 % of all cases; and on the other hand pancreatic neuroendocrine tumours (5 %), pancreatic acinar cell carcinoma (1-2 %), solid-pseudopapillary neoplasm (1-2 %) and pancreatoblastoma (<1 %) [170].

In 1979, Kozuka et al. [114] hypothesised that the progression towards carcinoma occurs via non-papillary hyperplasia through papillary and atypical hyperplasia. Nowadays, three different precursor lesions (pancreatic intraepithelial neoplasm (PanIN)) are distinguished: PanIN1, which is a flat (1A) or papillar (1B) hyperplastic lesion without dysplasia, PanIN2 showing moderate dysplasia and PanIN3, also termed carcinoma in situ with a high-grade dysplasia. This nomenclature was revised in 2015 (Baltimore Consensus). For diagnosis they suggest to differentiate between low-grade PanIN, the previous PanIN1 and high-grade PanIN, previously PanIN2 and PanIN3. The low-grade PanIN is suggested to be handled as being benign and unlikely to develop to PDAC, whereas high-grade PanIN is suggested to be seen as precursor lesion [22].

The linear progression model, through the various PanIN stages (Figure 9) was supported by various molecular findings like an accumulation of mutations with increased PanIN states. Most commonly, kirsten rat sarcoma viral oncogene homolog (KRAS) is found already in the early states, whereas cyclin dependent kinase inhibitor 2A (CDKN2A) mutations can be found more often in the moderate and high-grade states, and tumor protein p53 (TP53) and SMAD family member 4 (SMAD4) mutations mostly occur in the high-grade state [162]. Additionally, to those four 'driver' mutations, there are many rare mutations reported occurring in PDAC that show the heterogeneity of this cancer and might be used for further subtyping [220].

There are, however, findings that do not support the proposed linear progression. Firstly, PanIN1 lesions are considered as benign as they not necessarily progress to cancer. Secondly, the fact that most advanced lesions are found adjacent to PDAC, leading to the question whether those cells are precursors or whether the PDAC

cells made their environment changing to this kind of lesions [162]. This hypothesis goes in line with findings by Murphy et al. [157] that showed that PanIN3s carried a higher somatic mutation burden in comparison to PDAC, and both types have 80-90 % of their somatic variants in common. Another aspect which makes the molecular analysis challenging, is that mostly PanINs are extracted and analysed in patients that had gone under surgery because of pancreatic cancer, which means that even if the PanIN lesions and the cancer tissue are not adjacent it was not determined that they are molecularly independent. Especially, one study caused doubt for PanIN3 lesions and their proposed driver mutations as they found no SMAD4 and only a few TP53 mutations in PanIN3 lesions extracted from tissue without PDAC. Within the same study they also did not observe the same mutations in the neoplastic lesions and their corresponding carcinoma [90]. Additionally, the catastrophic effect of chromothripsis, a complete collapse of the genome, which can occur at each state and its potential influence should not be ignored [162].
All these findings are inconsistent with the classical PanIN model and cause the need of an analysis model that does not impose or assume any time relation a priori.

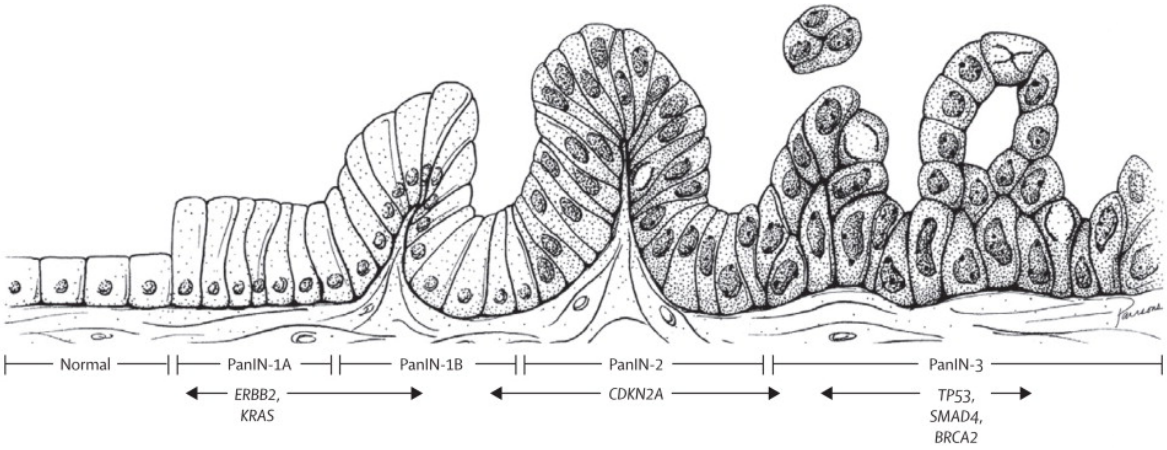


Figure 9: Evolution scheme of pancreatic adenocarcinoma (PDAC). PDAC is hypothesised to develop through a sequence of precursor lesions, pancreatic intraepithelial neoplasms (PanINs). Each stage is thought to be determined by mutations of specific genes. Early lesions (PanIN1A/ 1B) often carry mutations in erb-b2 receptor tyrosine kinase 2 (ERBB2) or kirsten rat sarcoma viral oncogene homolog (KRAS). PanIN2 is associated with cyclin dependent kinase inhibitor 2A (CDKN2A) mutations and PanIN3 with mutations in tumor protein p53 (TP53), SMAD family member 4 (SMAD4), BRCA2, DNA repair associated (BRCA2). Reprinted from [216], with permission from Elsevier and also credit to [93], with permission of the American Association for Cancer Research.

1.5.3 Colorectal cancer and hepatocellular carcinoma

Globally, colorectal cancer ranks fourth for cancer incidence and second for mortality, with approximately being the cause of 1 out of 10 cancer cases and deaths. Over the last decades it rose in incidence but declined in mortality, related to changes in life style, dietary and life-span elongation on the one hand and early detection and better treatment on the other hand [179, 31].
Starting in the 1980s the hypothesis of an adenoma-carcinoma sequence, describing the evolution of normal tissue via an adenoma towards a carcinoma, which might metastasise, developed based on morphological and epidemiological arguments [156, 76]. This hypothesis was later supported by molecular findings [64, 12]. In a very simplified way, todays point of view can be expressed as a multi-step progression with mutations in adenomatous polyposis coli (APC) leading to the transition of normal tissue to early adenoma and a subsequent alteration in KRAS. Mutations of TP53 might then cause the final transition from adenoma to adenocarcinoma [12]. As cancer is largely heterogeneous the actual situation is much more complex, as large amounts of various mutations are found investigating colorectal cancer [239].

Liver cancer is among the first four cancer types in terms of mortality and the sixth cancer type in terms of incidence. The two most common types of liver cancer are hepatocellular carcinoma (HCC) making up approximately 75-85 % of all cases and intrahepatic cholangiocarcinoma comprising 10-15 % [31]. Alterations in TP53, cell-cycle related genes and WNT-catenin beta 1 (CTNNB1), a β-catenin encoding gene, as well as in telomere maintenance, chromatin modifiers and inflammatory pathways are commonly found. HCC is proposed to develop through a sequence of precursor lesions, which includes inflammation stages [148]. Including chronic hepatitis as a potential precursor lesion is not surprising as an infection with hepatitis B virus (HBV) or hepatitis C virus (HCV) is among the main risk factors for HCC [31]. Hereby, the end point of a chronic liver disease is called cirrhosis, describing the replacement of normal tissue with fibrotic scar tissue. HCC is hence often proposed to develop via a cirrhotic stage [148].

1.6 Aims and overview

The aim of this thesis was to investigate ordinal phenotype relations within high-dimensional spaces in general and more specifically in pre-cancer stages.
To do so, I in collaboration with colleagues enhanced a mathematical and algorithmic procedure to estimate phenotype sequences from high-throughput data. In the following this method was characterised by me using artificial data. Furthermore, I investigated its power to reconstruct relational patterns and compared it to potential other methods. Afterwards I applied this newly proposed method to real data among them data of developmental stages and pre-cancer stages.
I employed an additional measure, network entropy, for the analysis of the hypotheses on ordinal relations of pre-cancer stages. This measure combined the information of the protein-protein network and the gene expression data. After having investigated the impact of correction methods for protein-protein networks on the network entropy, I applied the optimal workflow to the data of pre-cancer stages.

This thesis is structured into an introduction (Chapter 1), one part describing all methods and datasets applied (Chapter 2) and four result parts. The first result part comprises the advancements, further developments, the dimensionality dependency and the wrapping into an R-package of the Occ as well as its characterisation using artificial data (Chapter 3).
Within the second part, I concentrated on the type of order. I evaluated the type of ordinality detected by a binary classifier without disconnected decision regions (linear support vector machine (SVM)), compared it to one with disconnected decision regions (k-nearest neighbour (kNN)) and came up with four different criteria that might serve for characterising types of ordinality (Chapter 4) [120].
The third part comprises the screening results on real data (Chapter 5).
Finally, I evaluated the impact of errors in the protein-protein network and various correction methods on the entropy and analysed pre-cancer stages in terms of their network entropies (Chapter 6).

A conclusion, concluding the thesis and the major results (Chapter 7) as well as a short summary (Chapter 8) complete this thesis.

2

METHODS AND DATASETS

2.1 Notation

The following notation will be used throughout this thesis.
Vectors are written small and in bold face. I will use a sample set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{|\mathbf{X}|}$. The samples are given as n-dimensional feature vectors $\mathbf{x}_i = (x_1, \ldots, x_n)^T$, with $x_i \in \mathbb{R}$. Furthermore, I will call $\mathcal{Y} = \{y_i\}_{i=1}^{|\mathcal{Y}|}$ the label set, with $y_i \in \mathbb{N}$. If each sample in $\mathbf{X}$ is assigned to one category $y \in \mathcal{Y}$ resulting in a labelled sample $(\mathbf{x}, y)$, I will denote the whole labelled sample set as $\mathbf{X}_{\mathcal{Y}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{|\mathbf{X}_{\mathcal{Y}}|}$, $y_i \in \mathcal{Y}$ and the set comprising only samples of class y as $\mathbf{X}_y$. I will call the categories classes or phenotypes describing the semantic level and the samples, measurements describing the feature level. The symbol n describes the dimension, which refers to the number of features often referring to genes in gene expression data. The number of samples is m.

Furthermore, an ordinal semantic relation between labels y_i, y_j is denoted as $y_i \prec y_j$, with y_i being the predecessor and y_j being the descendent. If there exists one ordinal linear relation between all labels in $|\mathcal{Y}|$ it is called a total order or cascade, otherwise if an ordinal linear relation exists only between classes of a subset of classes, this relation is called a subcascade. A linear relation is meant as no class within the (sub)cascade having more than one predecessor or descendent.

A network spanned between a set of genes $g \in \mathbf{G}$ is denoted as gene-gene or protein-protein interaction (PPI) network. As I base the gene interaction networks on the interactions measured between their products, the proteins, both refer to the same network in my case. The network is given as a directed graph $\mathrm{PPI} = \{\mathbf{G}, \mathbf{E}\}$, with $\mathbf{G}$ being the set of genes given as vertices or nodes and $\mathbf{E} \subseteq \{g_i, g_j\} : (g_i, g_j) \in \mathbf{G}x\mathbf{G}, i < j$ being the set of edges or interactions. If a weight is assigned to each edge, the graph is called weighted and otherwise unweighted. The adjacency matrix A of this graph is denoted as A_{PPI}. All genes that share a direct edge with gene g are first-level neighbours of g and in the set $\mathbf{N}(g)$. Given a gene $g' \in \mathbf{N}(g)$, g' is called

descendent or child node if the edge starts at g and ends at g', otherwise g' is called predecessor or ancestor. The set of all first-level child nodes of g is denoted as $\mathbf{N}^{desc}(g)$ and the set of all first-level ancestor nodes is $\mathbf{N}^{anc}(g)$. Furthermore, $deg(g) = |\mathbf{N}(g)|$ is the degree of node g, with $deg^{out}(g) = |\mathbf{N}^{desc}(g)|$ and $deg^{in}(g) = |\mathbf{N}^{anc}(g)|$ being out-degree and in-degree, respectively.

2.2 Classification

With the possibility of generating high-throughput data, the need of efficient techniques that help to better understand the data, learn information from it and use the gained information for predictions on new data grew. Hereby, learning information is rather meant as eliciting general structural patterns than just storing data. Machine learning is a field that deals with the process of learning a general abstract model based on a set of measurements, which can later be used for predictions on new data.

The learning strategies can be divided in unsupervised and supervised methods, dependent on whether they are based on unlabelled input data $\mathbf{X}$ or on labelled input data $\mathbf{X}_y$. More explicitly supervised learning is characterised by a labelled training set and aims to find a rule that can be used to correctly assign labels to new samples, based on labelled examples. It can be used to tackle the problem of classification, which is the task of correctly categorising new data based on known data.

2.2.1 Learning a classifier

To learn and evaluate a classifier the labelled input set $\mathbf{X}_y$ is usually split into a training set $\mathbf{T} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{|\mathbf{T}|}$, $y_i \in \mathcal{Y}$ and a validation set $\mathbf{V} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{|\mathbf{V}|}$, $y_i \in \mathcal{Y}$, with $\mathbf{T} \cap \mathbf{V} = \emptyset$ and $\mathbf{T} \cup \mathbf{V} = \mathbf{X}_y$.
A classifier function $c : \mathcal{X} \to \mathcal{Y}$ is trained to assign one category $y \in \mathcal{Y}$ to an object $\mathbf{x}$ using the training data $\mathbf{T}$ [84]. It partitions the input space $\mathcal{X}$ into a set of labelled partitions $\{\mathcal{P}_1, \mathcal{P}_2, \ldots\}$. The union of all partitions of the same label y is called decision region $\mathcal{D}_y$. In this thesis $\mathcal{X} = \mathbb{R}^n$, if not explicitly stated otherwise.
In order to tune the parameters of the classifier cross-validation (CV) might be employed, for example in cases of small sample sizes (Section 2.2.5). The trained classifier is evaluated based on the validation set $\mathbf{V}$ and might be suitable to be used to classify new data.

2.2.2 Terminology

In general one distinguishes between binary classification tasks $|\mathcal{Y}| = 2$ and multi-class classification problems $|\mathcal{Y}| > 2$. In a binary classification task c_{y_i,y_j} will be written in short as c_{ij} and denotes the classifier, which was trained to distinguish between labels y_i, y_j, with $i \neq j$.
Within a multi-class setting $|\mathcal{Y}| > 2$ a classifier c might be trained on a subset of all classes $\mathcal{Y}' \subset \mathcal{Y}$. All classes y in the set of remaining classes $y \in \mathcal{Y} \setminus \mathcal{Y}'$ are considered as external or foreign classes for this classifier.

2.2.3 Performance measures

A classifier can be evaluated using different performance measures.

Accuracy
Given a label set $\mathcal{Y}$ and a labelled sample set $\mathbf{X}_{\mathcal{Y}}$ the accuracy of a classifier c is given as

$$\mathrm{acc}_c = \frac{1}{|\mathbf{X}_{\mathcal{Y}}|}\left(\sum_{\mathbf{x}\in\mathbf{X}_y} \mathbb{I}_{[c(\mathbf{x})=y]} + \sum_{\mathbf{x}\notin\mathbf{X}_y} \mathbb{I}_{[c(\mathbf{x})\neq y]}\right), \tag{1}$$

with $\mathbb{I}$ being the indicator function.

Conditional prediction rates
The general conditional prediction rate is

$$p_c(y|\mathbf{X}_{y'}) = \frac{1}{|\mathbf{X}_y|}\sum_{\mathbf{x}\in\mathbf{X}_y} \mathbb{I}_{[c(\mathbf{x})=y']}, \tag{2}$$

with $\mathbb{I}$ being the indicator function.

I will differentiate between three different kinds of conditional prediction rates

- sensitivity, if $y = y'$ and $y, y' \in \mathcal{Y}$

$$\mathrm{sens}_c(y) = p_c(y|\mathbf{X}_y) \tag{3}$$

- confusion, if $y \neq y'$ and $y, y' \in \mathcal{Y}$
- external rate, $y' \notin \mathcal{Y}$ and $y \in \mathcal{Y}$.

Purity
If a classifier c trained on samples of class $\mathcal{Y}$ is applied to samples of a class $y' \notin \mathcal{Y}$ the percentage of the most frequently predicted class label $y \in \mathcal{Y}$ for these samples $\mathbf{x} \in \mathbf{X}_{y'}$ is defined as the purity of y'

$$\text{pur}_c(y') = \max_{y \in \mathcal{Y}} \quad p_c(y \mid \mathbf{X}_{y'}). \tag{4}$$

The class y' is called external class [122].

2.2.4 Overfitting

During training the classifier is optimised on the given set of labelled data. If the performance is solely evaluated based on these data the classifier is prone to overfitting, which means that it has a good performance on the training data but a low performance on new unseen data. The performance on new unseen data is denoted as generalisation performance [84]. To estimate the generalisation performance the trained classifier is typically evaluated on an independent set of labelled validation data $\mathbf{V} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{|\mathbf{V}|}$, $y_i \in \mathcal{Y}$. If the validation set equals the training set it is called reclassification.

2.2.5 Cross-validation

In a perfect setting there would be a large set of labelled samples that can be split into two independent sets. If this cannot be done, for example due to a small sample set size CV can be used for model training and estimating the generalisation performance. In cross-validation the data is split into a fixed number of partitions (folds), each containing approximately the same number of samples. If additionally the class label distribution within the folds is similar to the one in the whole dataset the split is called stratified. For each fold training is performed on all data except the data of this fold and the data of this fold is used for evaluation. To account for the initial set separation these steps might be repeated several times. Finally, all performance estimates are averaged resulting in a final performance measure.
10 x 10 CV, which means splitting the data into 10 folds and repeating it 10 times, has become a standard procedure as 10 folds have to be shown experimentally to result in a good error estimate. Another possibility of splitting is leave-one-out-cross-validation (LOOCV), which trains on all samples except one. The sample that was left out is used for validation and the training-validation procedure is repeated for each sample [233]. LOOCV and 10 x 10 CV show different characteristics in terms of bias and variance. Whereas LOOCV has a low bias and a high variance, the variance of 10 x 10 CV is low but bias might be high [84].
I hence always used the same cross-validation method within one experimental

setup. I chose the method depending on the number of samples per class. If the number of samples per class is small for all classes within all datasets of an experimental setup I used LOOCV, otherwise I used 10 x 10 CV.

2.2.6 Binary classifier systems

Binary classifier systems deal with binary classification tasks ($|\mathcal{Y}| = 2$), splitting the input space into exactly two decision regions. In the following I describe the binary classifiers used throughout this thesis.

2.2.6.1 Support vector machine

The support vector machine (SVM) belongs to the group of linear classifiers [50, 214, 2].
Linear classifiers separate two groups of classes based on a linear hyperplane $\mathbf{w}^T\mathbf{x} - b = 0$. Here $\mathbf{w} \in \mathbb{R}^n$ describes the multidimensional angle und $b \in \mathbb{R}$ denotes the threshold describing the shift of the hyperplane with respect to the origin. A linear classifier is trained on a given training set $\mathbf{T}$ but there might exist several equally performing classifiers on that set. One, which imposes a further constraint, is the SVM.
It is a large margin classifier, which means that this method aims in finding one separating hyperplane such that the area, that contains no samples around this hyperplane is as large as possible (Figure 10A). The samples that have the minimal distance towards this hyperplane are called support vectors.

Given a set of labelled training data $\mathbf{T} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{|\mathbf{T}|}$ and $\mathcal{Y} = \{1, -1\}$ the constrained optimisation problem, resulting in the hyperplane parameters $\mathbf{w}$ and b, is

$$\begin{aligned} &\underset{\mathbf{w},b}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{w}\|_r^2 \text{ with,} \\ &y_i(\mathbf{w}^T\mathbf{x_i} + b) \geqslant 1, \text{ for all } 1 \geqslant i \geqslant |\mathbf{T}|. \end{aligned} \tag{5}$$

If $r = 1$ a type 1 regularisation is performed, which shows an intrinsic feature selection (L1-SVM) otherwise if $r = 2$ one deals with a type 2 regularisation (L2-SVM). I use $r = 2$ if not stated otherwise.

The classifier function is

$$c(\mathbf{x}) = \begin{cases} 1 & \text{if} \quad \mathbf{w}^T\mathbf{x} \geqslant b \\ -1 & \text{else.} \end{cases} \tag{6}$$

Using the nonnegative Lagrange multipliers $\lambda = (\lambda_1, ..., \lambda_m)^T$ the constrained problem can be converted to an unconstrained problem in its primal form

$$Q(\mathbf{w}, b, \lambda) = \frac{1}{2}\|\mathbf{w}\|_r^2 - \sum_{i=1}^{|\mathbf{T}|} \lambda_i (y_i(\mathbf{w}^T \mathbf{x_i}) + b - 1), \tag{7}$$

which is maximised for $\lambda \geqslant 0$ and minimised with respect to $\mathbf{w}$ and b.

The previously described hard-margin SVM only finds a solution if the data is linear separable. To allow for a certain amount of misclassifications a loss term is introduced resulting in a soft-margin SVM

$$\begin{aligned} &\underset{\mathbf{w},b,\xi}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{w}\|_r^2 + C\sum_{i=1}^{m} \xi_i \text{ with,} \\ &y_i(\mathbf{w}^T \mathbf{x_i} + b) \geqslant 1 - \xi_i, \text{ for all } 1 \geqslant i \geqslant m. \end{aligned} \tag{8}$$

The loss term consists of two further variables, the slack variable ξ, quantifying the amount of correct classification violation, and the cost parameter C, which defines the stress of the slack variable [37]. The cost parameter is handled as hyperparameter, not included in the optimisation process. For high-dimensional data it was shown that the solution is independent of the chosen cost parameter [160].

Dealing with non-linear separable problems the kernel trick can be applied. By mapping the input vector $\mathbf{x}$ to higher dimensions linear separability is enhanced. However instead of using a mapping function $f(\mathbf{x})$ directly, a kernel function $f_k(\mathbf{x}, \mathbf{x}') = f^T(\mathbf{x})f(\mathbf{x}')$ is used, which makes it possible not to handle the high-dimensional space explicitly. Kernels that are used within this study are:

- linear kernel: $f_k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T\mathbf{x}'$
- polynomial kernels: $f_k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T\mathbf{x}' + 1)^q$, with degree q.

2.2.6.2 k-nearest neighbour

The k-nearest neighbour (kNN) classifies new samples based on the number of labelled samples in its neighbourhood [66](Figure 10B). Hereby, the distances $\mathbf{D_v}$ between the new sample $\mathbf{v}$ and all labelled samples $(\mathbf{x}, y) \in \mathbf{T}$ are calculated. Next, the k-nearest neighbourhood is defined as the samples $(\mathbf{x}, y) \in \mathbf{T}$ that are ranked (rk denotes the rank function) as k-nearest within the amount of distances $d \in \mathbf{D_v}$.

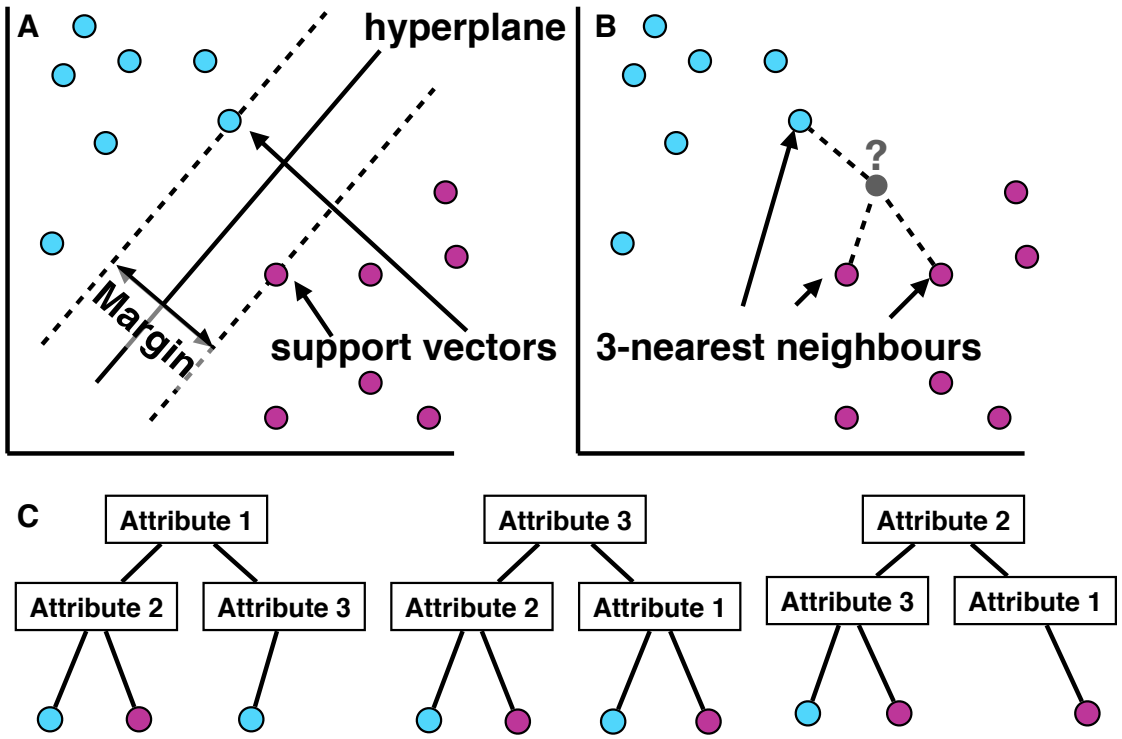


Figure 10: Three types of binary classifier systems. The samples of two different classes are indicated as blue and purple dots. A: The support vector machine (SVM) finds the hyperplane that separates the two classes such that the distance between the hyperplane and the support vectors is maximised. The support vectors are those samples that are closest to the hyperplane. B: shows an example of a 3-nearest neighbour. The new sample is evaluated based on the label of the 3 samples of the labelled training data that are closest to it. C: The random forest (rF) is an ensemble of decision trees. The decision at each node is based on an attribute and the composition and order of attributes differs among the trees. An unknown sample is labelled based on a majority vote of these trees.

Last a majority vote of the labels y of the k-nearest points NN_k is performed to result in the classification result $c(\mathbf{v})$ of sample v.

$$c(\mathbf{v}) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} |\{(\mathbf{x}, y) \in NN_k(\mathbf{v}, \mathbf{T})\}| \tag{9}$$

$$NN_k(\mathbf{v}, \mathbf{T}) = \{(\mathbf{x}, y) \in \mathbf{T} | rk_{\mathbf{D_v}}(d(\mathbf{v}, \mathbf{x})) \leqslant k\} \tag{10}$$

$$\mathbf{D_v} = \{d(\mathbf{v}, \mathbf{x}) | (\mathbf{x}, y) \in \mathbf{T}\} \tag{11}$$

I used the Euclidian distance as distance measure and performed a hyperparameter tuning for k, with $k \in \{1, 3, 5\}$, if not stated otherwise.

2.2.6.3 Random forest

The random forest (rF) is an ensemble of several tree predictors (Figure 10C) [33, 32], with each tree being trained on a set of samples. The samples are drawn randomly with replacement from the set of all samples. The results of the individual trees are fused using majority vote.
The decision tree model itself is based on a tree like topology. Each internal node represents a feature and the two arcs towards their child nodes represent different value ranges for that feature. The leaves represent the classes. One tree is usually built up by using the feature that best splits the data as next node. To measure the goodness of split the information gain or the Gini impurity is used [33]. Within the scope of this thesis Gini impurity was used. Neither of the two measures was shown to be more accurate and Gini impurity is computationally less expensive. It measures how likely a sample is misclassified if the label is randomly based on the distribution of class labels within the training set.

2.2.7 Multi-class classification

Multi-class classification deals with classification tasks of more than two classes ($|\mathcal{Y}| > 2$). There are two general strategies to deal with this: on the one hand there exist classifier systems, that easily adapt to more than two classes, like kNN; On the other hand the task can also be decomposed into binary tasks, being solved using binary classifier systems and fusing the individual results afterwards. One decomposing strategy is *one-against-one* (OaO).

2.2.7.1 One-against-one

Given a set of classes $\mathcal{Y}$ the OaO [27] or *pairwise* approach trains one binary classifier c_{ij} for all pairs of labels $y_i, y_j \in \mathcal{Y}$, with $i < j$ resulting in an ensemble $\mathcal{E}_{OaO}$ of $(|\mathcal{Y}|(|\mathcal{Y}|-1)/2$ binary classifiers. The pairwise decisions are combined using a majority vote. This means the label $y \in \mathcal{Y}$ that is predicted by the most pairwise trained classifiers is assigned to the sample $\mathbf{x}$. The classifier function $h_{OaO} : \mathcal{X} \to \mathcal{Y}$ for samples $\mathbf{x}$ is expressed as

$$h_{OaO}(\mathbf{x}) = \arg\max_{y \in \mathcal{Y}} \sum_{c \in \mathcal{E}_{OaO}} \mathbb{I}_{[c(\mathbf{x})=y]} \quad \text{with} \quad \mathcal{E}_{OaO} = \{c_{ij} | i < j \leqslant |\mathcal{Y}|\}_{i=1}^{|\mathcal{Y}|-1}, \tag{12}$$

with $\mathbb{I}$ being the indicator function. If the majority is not distinct the label is drawn randomly from the set of most often predicted labels.

2.2.8 Ordinal classification

In a classical ordinal classification task a known order between k class labels ($y_1 \prec ... \prec y_k$) is used to further constrain the classifier and improve classification performance. Hereby, the order on the labels is only assumed to be reflected in the investigated data. An assumption, which is usually not questioned but does not necessarily have to be true.
If I refer to the position of a class in an order I will use parenthesised indices. As consequence, I will denote semantic relations between class labels y_i, y_j as $y_i \prec y_j$, which means class y_j is semantic ordinal related to y_i, in contrast to $y_{(i)} \prec y_{(j)}$, which translates as the class at position j is semantic ordinal related to the class at position i in a given order.

2.2.8.1 Ordinal classifier cascade

The ordinal classifier cascade (Occ) belongs to the group of ordinal binary decomposition methods. It was introduced by Lattke* et al. [119], who evaluated its performance for different training schemes and types of base classifier. It was shown that the performance of the Occ is strongly dependent on the given order and best distinguishes the correct order from the others for all base classifiers if trained in a pairwise manner.

The Occ is an ensemble multi-class classifier system, consisting of a set of $|\mathcal{Y}| - 1$ pairwise trained base classifiers $\mathcal{E}_{Occ} = \{c_{(i)}\}_{i=1}^{|\mathcal{Y}|-1}$ for a given set of class labels $\mathcal{Y}$. The classifier $c_{(i)} : \mathcal{X} \to \{y_{(i)}, y_{(i+1)}\}$ is trained to distinguish between class $y_{(i)}$ and $y_{(i+1)}$ of a given semantic order $y_{(1)} \prec ... \prec y_{(|\mathcal{Y}|)}$. It denotes the classifier at the ith position of an Occ.
Given, additionally, samples $\mathbf{x} \in \mathbf{X}$, the ordinal classifier cascade is a function $h_{Occ} : \mathcal{X} \to \mathcal{Y}$ such that

$$h_{Occ}(\mathbf{x}) = \begin{cases} y_{(i)} & \text{if } c_{(i)}(\mathbf{x}) = y_{(i)} \wedge \forall j < i : c_{(j)}(\mathbf{x}) \neq y_{(j)} \\ y_{(|\mathcal{Y}|)} & \text{else.} \end{cases} \tag{13}$$

The procedure of this cascades works by evaluating one binary classifier after the other. This means that a sample $\mathbf{x}$ is first evaluated by the classifier $c_{(1)}$. If this classifier predicts its first class $y_{(1)}$ the cascade stops and the label is assigned to $\mathbf{x}$. If the classifier predicts its second class $y_{(2)}$, the sample is transferred to the second binary classifier $c_{(2)}$. This goes until the sample is either labelled or the end of the cascade is reached. If the last binary classifier predicts its second class the sample

is classified as the last class of the cascade. Its functionality is depicted in Figure 11. As a performance measure the minimal class-wise sensitivity

$$\text{smin} = \min_{y \in \mathcal{Y}} \text{sens}_c(y) \tag{14}$$

with $\text{sens}_c(y)$ (Equation 3) is used.

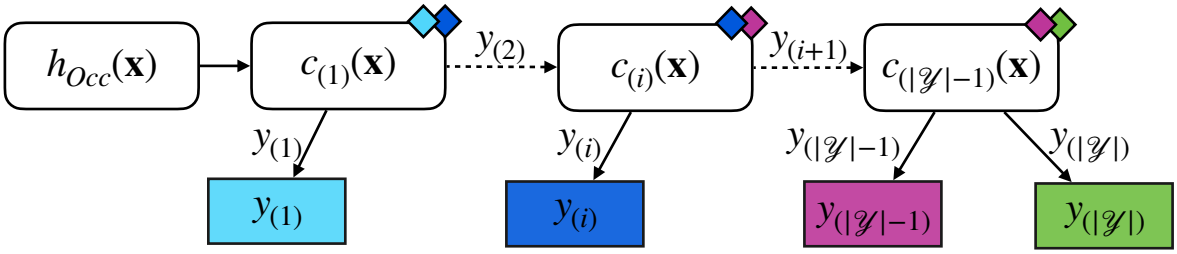


Figure 11: Skeleton of the ordinal classifier cascade. The base classifiers are trained pairwise (labels given by coloured squares) and ordered based on a presumed order of the class labels. A sample is classified by evaluating the base classifiers sequentially. If one classifier predicts the sample as its first class the cascade stops and the sample is classified. Otherwise it is traversed to the next base classifier. If no base classifier predicts the sample as its first class the sample is labelled based on the last label within the sequence. See [123] licensed under CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

A four class example that shows the decision regions assignment at each step of an Occ using independently trained binary linear classifiers and assuming three different class orderings is given in Figure 12. As the decision regions differ between the correct and the reverse order (Figure 12 A,B), some samples might not be classified correctly assuming a reversed order, resulting in a diminished performance. This illustrates that the Occ might be able to distinguish between forward and backward directions.

2.2.8.2 parallel SVM

Binary base classifiers might also be trained dependent on each other. Using the SVM as base classifier the multidimensional angle $\mathbf{w}$ can be constrained to be the same for all base classifiers [120]. In this case they do only differ in their thresholds and the boundaries are parallel. This construct, however, requires a simultaneous training of all binary classifiers

$$\forall i = 1, \ldots, |y| - 1 : c_{(i)}(\mathbf{x}) = \begin{cases} y_{(i)} & \text{if} \quad \mathbf{w}^T\mathbf{x} \geqslant b_i \\ y_{(i+1)} & \text{else} \end{cases} . \tag{15}$$

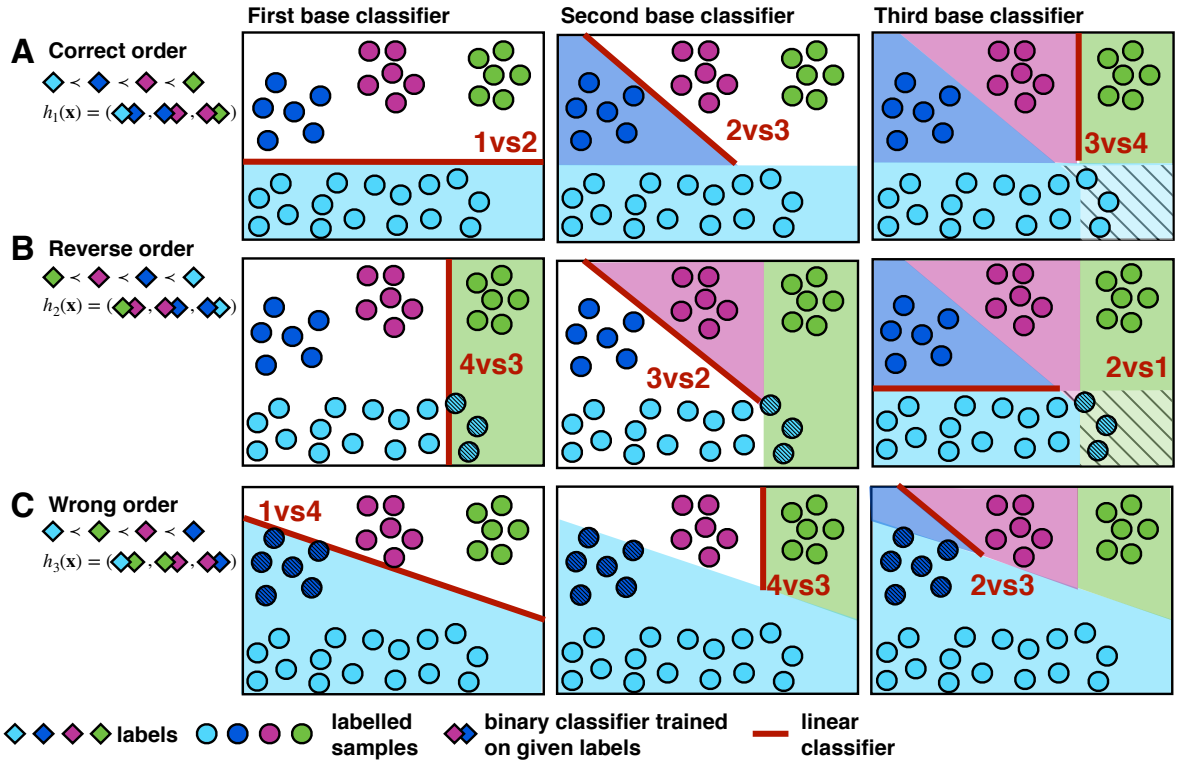


Figure 12: Example of decision region assignment using an ordinal classifier cascade. Given four classes (represented by the different colours) and several samples located in a 2D features space, the samples are classified guided by the assumed order. Each coloured circle within the feature space represents a sample. Wrongly classified samples are marked with stripes. There are three order assumptions given: correct order, reverse order and wrong order. The corresponding assumed label orders are depicted below each title. The assumed order defines the order of the base classifiers in each cascade $h(\mathbf{x})$. A: If the order is reflected in the feature space the samples are assigned to their correct classes. B: Depending on the location of the samples in the feature space, there might be a difference in the performance of the correct and its reverse order. The region that is responsible for distinguishing the difference between correct and reverse order is marked with stripes in the scheme of the third base classifier in A and B. C: If a wrong order is assumed many samples are classified wrong and the performance of the ordinal classifier cascade in terms of minimal class-wise sensitivity drops.

2.2.8.3 Result evaluation

Ordinal classification can be performed on the whole set $\mathcal{Y}$ or on subsets. If only a subset is used the size of this subset might be included in evaluating the performance of a classifier, resulting in a multi-objective optimisation problem.

Pareto optimal solution

Multi-objective optimisation deals with the problem of having more than one objective that is to be optimised. The Pareto optimal set contains all solutions that cannot be further optimised without reducing the performance of another objective. Searching for ordinal relations consisting of well-distinguishable classes, there might be the objective of the longest relations that perform best. In this setting the Pareto optimal set might contain relations of different lengths with shorter lengths showing a better performance (Figure 13) [62].

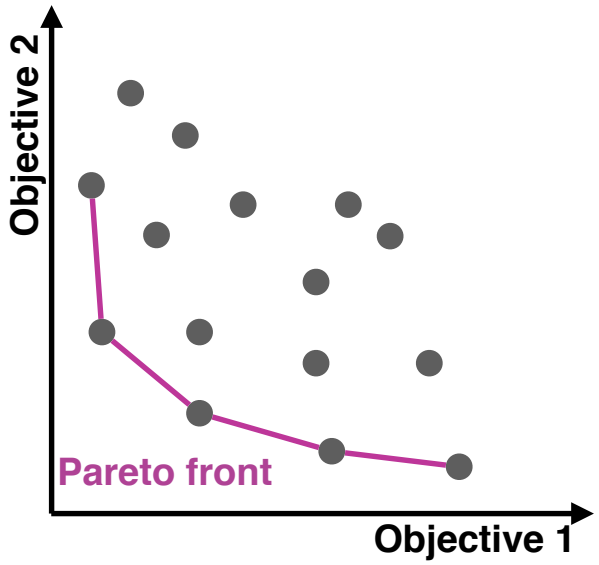


Figure 13: Pareto optimal solutions. Depicted is a multi-objective optimisation problem with two objectives that both are to be maximised. The grey dots show possible solutions. The Pareto optimal set comprises all solutions (dots connected by a purple line) that cannot be further optimised without the reduction of one of the two objectives.

2.2.9 Screening procedure

I performed 10 x 10 CV or LOOCV experiments for all trainings and tests of one dataset collection. The cross-validation procedure was chosen depending on the dataset sample sizes and resulted in LOOCV for the artificial and cancer dataset screenings. The fold lists were generated once in a stratified way and always the same fold list was used for a dataset. I used the *TunePareto* R-package [158] for training and testing. This package includes the SVM from *e1071* [151]. If not stated otherwise, the default cost parameter of 1 was kept, as it was shown that this parameter has a negligible impact in high dimensions [160]. The experiments with the L1-SVM were performed using the R-package *LiblineaR* [88]. The kNN and

random forest (rF) function is taken from *class* and *randomForest* R-packages [215, 129], respectively. The hyperparameters are given at each experiment.

2.3 Reconstructing neighbourhood embeddings

Distances between samples of different classes might be used to imply neighbourhoods between class labels. Under the assumption that labels of close samples are more semantic ordinal related than labels of samples showing a high distance, these neighbourhoods might be used to reconstruct orders on the labels. Consequently, these orders are considered to be embedded and reflected in the feature space.
Another strategy might be the embedding of a predefined topology to the data and the evaluation of the goodness of this embedding. In this setting, the topology defines the neighbourhood structure and possible relations on the labels.
Network reconstruction describes this procedure of eliciting a neighbourhood network from data.

2.3.1 Dependency measures

Dependency measures might be used to estimate the strength of relations. Two of them are correlation based methods [197] and methods based on mutual information [254]. I used the following measures given two samples $(\mathbf{x}, \mathbf{x}') \in \mathbf{X}$, their means $\mu_{\mathbf{x}} = \frac{\sum_{i=1}^{n} x_i}{n}$ and $\mu_{\mathbf{x}'} = \frac{\sum_{i=1}^{n} x'_i}{n}$, standard deviations $\sigma_{\mathbf{x}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu_x)^2}{n-1}}$ and $\sigma_{\mathbf{x}'} = \sqrt{\frac{\sum_{i=1}^{n}(x'_i - \mu_{\mathbf{x}'})^2}{n-1}}$, and their covariance $cov(\mathbf{x}, \mathbf{x}') = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{\mathbf{x}}) \cdot (x'_i - \mu_{\mathbf{x}'})$.

Correlation
The correlation measures are:

- Pearson correlation coefficient [169]: $\rho_p = \frac{cov(\mathbf{x},\mathbf{x}')}{\sigma_{\mathbf{x}}\sigma_{\mathbf{x}'}}$.
- Spearman's rank correlation coefficient [196]: $\rho_s = \frac{cov(rk(\mathbf{x}),rk(\mathbf{x}'))}{\sigma_{rk(\mathbf{x})}\sigma_{rk(\mathbf{x}')}}$, with rk as rank function $rk : \mathbb{R}^n \to \mathbb{N}^n, rk(\mathbf{x}) = \sum_{i=1}^{n} \mathbb{I}(\mathbf{x} - x_i)$ and $\mathbb{I}$ being the indicator function.
- Kendall rank correlation coefficient [108]: $\rho_k = \frac{N_{con} - N_{dis}}{\binom{n}{2}}$, with N_{con} being the number of concordant pairs and N_{dis} being the number of discordant pairs. A pair (x_i, x'_i) and (x_j, x'_j) is concordant if they are of the same order ($x_i < x_j$ and $x'_i < x'_j$ or $x_i > x_j$ and $x'_i > x'_j$), otherwise it is discordant.

Mutual information

The mutual information (MI) is a measure derived from information theory. For two variables $(\mathbf{x}, \mathbf{x}') \in \mathbf{X}$, with $\mathbf{x} = (x_1, \ldots, x_n)^T$ and $\mathbf{x}' = (x'_1, \ldots, x'_n)^T$, it is defined as

$$MI(\mathbf{x}, \mathbf{x}') = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{x}'), \tag{16}$$

and $H(\mathbf{x}) = -\sum_{i=1}^{n} prob(x_i) \log(prob(x_i))$, being the entropy and $H(\mathbf{x}|\mathbf{x}') = -\sum_{i=1}^{n} prob(x_i|x'_i) \log(prob(x_i|x'_i))$ the conditional entropy, respectively. $prob$ denotes a probability distribution.

If the probability distribution is unknown it can be estimated using the empirical probability distribution $\widetilde{prob}(\mathbf{x})$ of a sample $\mathbf{x}$. Hereby, the sample $\mathbf{x}$ is discretised into $\tilde{m}$ equally-wide intervals and the number of events $\tilde{x}$ falling into each interval is calculated. For each interval $i \in (1, \ldots, \tilde{m})$ the relative frequency of events per interval $f_{rel}(x_i) = \frac{\tilde{x}_i}{\sum_j^{\tilde{m}} \tilde{x}_j}$ is used as probability estimate $\widetilde{prob}(x_i)$. The conditional probability can be estimated using the joint probability distribution $\widetilde{prob}(\mathbf{x}, \mathbf{x}')$ as $\widetilde{prob}(\mathbf{x}|\mathbf{x}') = \frac{\widetilde{prob}(\mathbf{x}, \mathbf{x}')}{\widetilde{prob}(\mathbf{x}')}$.

Using the empirical probability distribution the entropy H_{emp} is calculated and used for MI calculation.

I use a normalised MI $NMI(\mathbf{x}, \mathbf{x}') = \frac{2 \cdot MI(\mathbf{x}, \mathbf{x}')}{H(\mathbf{x}) + H(\mathbf{x}')}$ [201] and the following two entropy estimation methods:

- empirical entropy [166]: $H_{emp} = -\sum_{i=1}^{n} \widetilde{prob}(x_i) \log(\widetilde{prob}(x_i))$, with $\widetilde{prob}$ being the empirical probability distribution (Equation 16).
- Miller-Madow corrected entropy [154]: $H_{MM} = H_{emp} + \frac{\hat{m}-1}{2 \cdot n}$ is a bias corrected calculation based on the empirical probability distribution, with $\hat{m}$ as number of bins with nonzero probability. It corrects for the bias caused by the larger error because of an underestimation on $\widetilde{prob}$ in comparison to an overestimation of $\widetilde{prob}$.

To evaluate the dependency between two sets of samples, with more than 1 sample per set, I used the average of the calculated sample-pairwise dependencies.

2.3.2 Self-organising maps

A technique that tries to embed a predefined topology is self-organising maps (SOMs) [112]. Using a linear topology SOMs might be suitable for finding linear relations between sample clouds. SOMs originated in the field of unsupervised learning. The idea is to fit a topological map to a given data and by this elicit the

topological structure of the data.
They are made up of a topological map TX_{map}, which consists of a set of units $\{TX_{map_j}\}_{j=1}^{a}$ that are connected based on a predefined topology (hexagonal/ rectangular grid). Each unit is described by a weight vector $\mathbf{w}_{TX_{map_j}}$ and the position on the map. I assume one unit TX_{map_j} per class $y_j \in \mathcal{Y}$.

SOMs are trained based on a sample set $\mathbf{X}_{\mathcal{Y}}$ in an iterative manner, starting at $u = 0$ with a random initialisation of all weight vectors $\mathbf{w}_{TX_{map_j}}$. During the training the distance $d^u(\mathbf{x}_i, j) = d^u(\mathbf{x}_i^j, TX_{map_j})$ between each unit ($j = (1, \ldots, a)$) and one sample $\mathbf{x}_i = (x_1, \ldots, x_n)^T$ is calculated. The winning unit $TX_{map_{j'}}$ is given by the minimal distance $\text{argmin}_j\, d(\mathbf{x}_i, j)$. The weights of the winning unit are updated to be closer to the sample $\mathbf{x}_i$, as $\mathbf{w}^{u+1}_{TX_{map_{j'}}} = \mathbf{w}^u_{TX_{map_{j'}}} + \epsilon^u \cdot (\mathbf{x}_i - \mathbf{w}^u_{TX_{map_{j'}}})$, with ϵ^u being the learning rate at step u. Additionally to the winning unit, a set of its topological neighbours is updated as well. The set comprises all topological neighbours that are closer to the winning unit than a given range τ^u. Both ϵ^u and τ^u decrease with the number of iteration steps.
This procedure is repeated iteratively for all samples until the solution converges or a predefined number of iteration steps is reached [112]. I used SOMs with a linear topology and afterwards assigned labels to the units based on a majority vote based on the labels of the samples that belong to the unit. The neighbourhoods of the units defined the final cascade of labels. If two neighboured nodes were assigned the same label I merged them to one category within the cascade.

Additionally, there exist supervised versions of SOMs, which also take the label information into account. In the same way as SOMs are closely related to vector quantisation, supervised versions of SOMs might be considered related to learning vector quantisation [112], which take the accordance of class labels between sample and the winning unit into account while updating the weights. Supervised versions of SOMs use a second map TY_{map} that has as many units as TX_{map} and each unit is assigned a label. One of these methods is the xy-fused method [150]. Here a combined distance is calculated combining for each sample $(\mathbf{x}, y)$ the distances $d(\mathbf{x}_i, TX_{map_j})$ and $d(y_i, TY_{map_j})$ on TX_{map} and TY_{map}. This combined distance $d(\mathbf{x}_i, j) = \kappa \cdot d(\mathbf{x}_i, TX_{map_j}) + (1 - \kappa) \cdot d(y_i, TY_{map_j})$ is used to define the winning unit. The parameter κ is a weighting parameter. I used an equal weighting scheme ($\kappa = 0.5$).

For the SOM screenings I used the R-package *kohonen* with the default settings if not stated otherwise [226, 227].

2.4 Network entropy

Network entropy, also termed signalling entropy, is a measure that incorporates the information of the gene measurements $\mathbf{x}$ with the information given by a gene interaction network $PPI = \{\mathbf{G}, \mathbf{E}\}$, with $\mathbf{G}$ being the set of vertices and $\mathbf{E}$ being the set of interactions [18]. Furthermore, one assumes that a sample $\mathbf{x}$ resulted from a stochastic process, like a Markov Process. This process is characterised by its transition matrix P, with p_{ij} describing the probability of a transition from gene g_i to g_j, with the expression values given by x_i and x_j. This probability p_{ij} functions as an interaction probability estimate. If there is no edge in $\mathbf{E}$ between g_i and g_j, $p_{ij} = 0$.
Otherwise it is estimated based on the mass action principle from the expression values x_i and x_j using the expressions' product [18]

$$p_{ij} = \frac{E_{ij}}{\sum_{k \in \mathbf{N}(i)} E_{ik}} \text{ , with } E_{ij} = x_i \cdot x_j \text{ , and } \sum_{j \in \mathbf{N}(i)} p_{ij} = 1. \tag{17}$$

With $\pi = \pi P$ being the stationary distribution the global network entropy rate is calculated as

$$SR = -\sum_{i,j}^{n} \pi_i \cdot p_{ij} \log(p_{ij}). \tag{18}$$

The maximal entropy of a network is the topological entropy introduced by Adler et al. [6] and solely based on the structure of the Markov Process given in the adjacency matrix A_{PPI}.

To ensure the existence of a stationary distribution the largest strongly connected component of PPI is used for the calculation of the global network entropy rate [145]. This refers to the largest subnetwork, in which each node can be reached by each other node in a directed network.

Network entropy can be seen as a measure of the complexity of a system. Interpreted as a signalling network, it can be said that the network entropy rate describes how determined the network flow is. In a network, in which each edge is chosen with the same probability the entropy rate is maximal. Whereas in a network, in which specific edges have higher probabilities the entropy decreases as there is more information about how flows in this network behave. Consequently, if there are various possibilities for an arriving signal to be processed in a cell, the entropy is higher than if it is known that one specific processing path is preferred as response.

A normalised local network entropy exists, too. It describes how determined the signalling flow is, that passes one specific gene. The measure is calculated per gene g_i as

$$S_i = -\frac{1}{\log(deg^{out}(g_i))} \cdot \sum_j p_{ij} \log(p_{ij}), \tag{19}$$

with deg^{out} being the out-degree of node g_i. The normalisation becomes necessary, because otherwise the entropy would automatically increase with a higher out-degree. Instead of the global network entropy rate, also the averaged local entropies could be used as a 'global' measure, however this would correspond to a non-equilibrium state [18].

2.4.1 Protein interaction networks

Gene interaction networks are based on the interactions measured between their products, the proteins.

2.4.1.1 Resources

I downloaded interactions from three different protein interaction databases and filtered the interactions as denoted:

1. Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) (version 10.5; 16.01.2018) [200]: filtered for only experiments and databases, 726174 interactions and 14913 proteins.
2. Biological General Repository for Interaction Datasets (BioGRID) (3.4.153; 25.09.2017) [164, 41]: only direct interaction, physical association, association, 354999 interactions and 20980 proteins.
3. Pathway Commons (PC) (version 9; 30.6.2017) [39]: all: 971618 interactions and 19589 proteins.

An overlap of the proteins and interactions is given in Figure 14A. Additionally, these characteristics are denoted for the strongly connected component (scc) extracted from these networks as well as from their union Figure 14B. This shows that each database resource contributes interactions not included in another database resource. If tested whether the networks and their components are scale-free a Kolmogorov-Smirnov test did not reject the hypothesis that the degree distribution could have been drawn from the fitted power-law distribution for all individual networks as well as the union.
The union of these networks was used as basis for the calculation of the network entropy on human data.

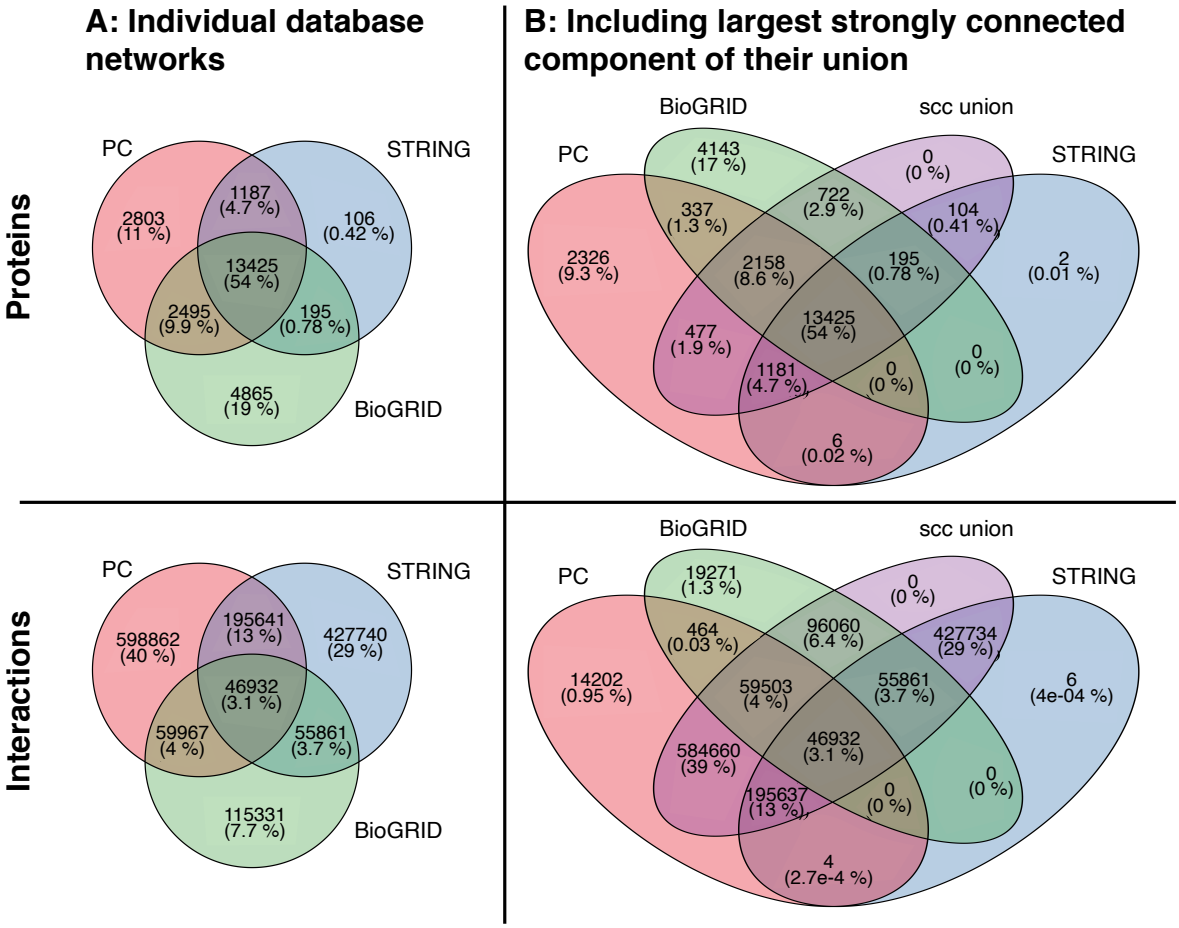


Figure 14: Venn diagrams of protein and interaction number of three database resources individually and including the extracted largest strongly connected component of their union. The number of proteins (upper part) and interactions (lower part) of A: the individual database networks and B: the extracted largest strongly connected component (scc) of the union in comparison to the original databases is shown. Given are the absolute numbers and the percentage. The Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) (version 10.5; 16.01.2018) was filtered for data that was experimentally confirmed or extracted from other databases. In case of the Biological General Repository for Interaction Datasets (BioGRID) database (3.4.153; 25.09.2017) the interactions were filtered for labelled as direct interaction, physical association, association. For Pathway Commons (PC) (version 9; 30.6.2017) no previous filtering was applied. Union comprises the scc of the union of all three databases.

2.4.2 False positive reduction strategies

Protein interaction networks are reported to have many false positive (FP) and false negative (FN) edges [141, 29]. They result from erroneous experimental as well as computational approaches. FN edges mean that there was no interaction

measured although an interaction exists. Correcting FNs would result in predicting interactions, which is an intensively studied field [58, 83] but out of the scope of this thesis. FP edges refer to false measured interactions.

There exists several methods that might be used to correct for FP edges within a PPI [99, 141, 35, 4, 184, 130, 187, 101, 225]. In the following three different categories are distinguished: Measures that rely purely on information revealed from the topology of the network under investigation, measures that include external semantic knowledge taken from gene ontology (GO) and measures that integrate several additional resource information. All of them are used in a similar fashion (Figure 15): Based on additional information sources, for example network topology or gene ontology, a score matrix R is calculated estimating how reliable an interaction between all pairs of genes is. Next, a threshold on the edge weights is imposed, all edges that do not pass this threshold are filtered out and the remaining edges are assigned 1 as value, leading to the correction matrix R'. In case of a soft thresholding no edges are filtered out and the reliability values are used as such. Finally, the corrected adjacency matrix is calculated as

$$A'_{PPI} = R' \circ A_{PPI}, \tag{20}$$

with $\circ$ being the Hadamard-product.

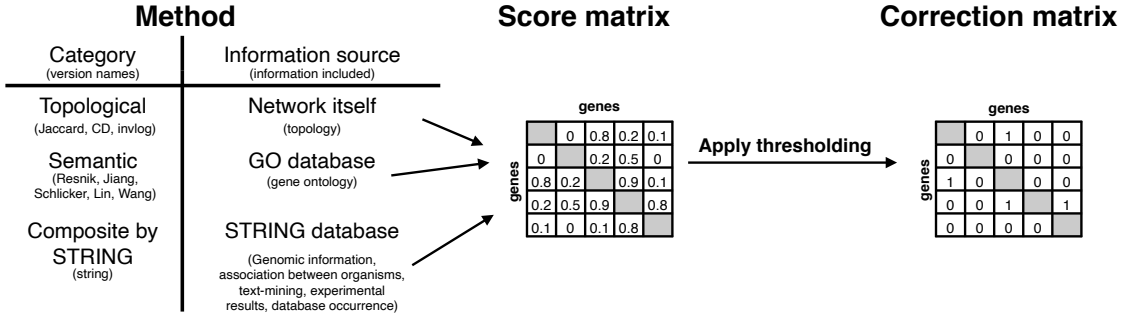


Figure 15: Scheme of the applied techniques to correct for false positive edges in a protein-protein interaction (PPI) network. For each method (topological (3 versions), semantic (5 versions) and composite by Search Tool for the Retrieval of Interacting Genes/Proteins (STRING)) first a scoring matrix is calculated using the material of the information source. A thresholding strategy is applied to these scores, resulting in the final correction matrix that is used for correcting the network under consideration. gene ontology (GO), Czekanowski-Dice (CD), inverse log-weighted (invlog).

2.4.2.1 Topology based scoring

The idea of using topological information of a protein-protein interaction network PPI is that two proteins interact more likely if their neighbours also interact. Given

is a $PPI = \{\mathbf{G}, \mathbf{E}\}$, with $\mathbf{G}$ being the set of vertices and $\mathbf{E}$ being the set of interactions. I use three different measures to estimate reliability of an interaction between a vertex g and g':

- Jaccard similarity [99, 141]: $R_J^{top}(g, g') = \frac{|\mathbf{N}(g) \cap \mathbf{N}(g')|}{|\mathbf{N}(g) \cup \mathbf{N}(g')|}$.
- The Czekanowski-Dice (CD) similarity is defined as $R_{CD}^{top}(g, g') = \frac{\mathbf{N}(g) \triangle \mathbf{N}(g')}{|\mathbf{N}(g) \cup \mathbf{N}(g')| + |\mathbf{N}(g) \cap \mathbf{N}(g')|}$. The symbol $\triangle$ describes the symmetric set, that are those elements that belong either to $\mathbf{N}(g)$ or to $\mathbf{N}(g')$ but not to both [35].
- The inverse log-weighted (invlog) similarity $R_{invlog}^{top}(g, g') = \sum_{g_n \in (\mathbf{N}(g) \cup \mathbf{N}(g'))} \frac{1}{\log(deg(g_n))}$ takes also the degree of the common neighbours into account, as it might be that interactions with high degree nodes occur more likely by chance [4].

The calculation is performed with the R-package *igraph* [53].

2.4.2.2 Semantic based scoring

I used five different methods, which use semantic information extracted from a taxonomy. Four of them are information content-based [184, 130, 187, 101] and one uses additional graph information [225].
I used GO [13, 207, 208] with the taxonomy of the compartments, as it is unlikely that two proteins interact if they are not located in the same compartment. Taxonomies are used for categorisation and this taxonomy has a hierarchical structure made up of terms and with "cellular_component" as root term. Hereby, not all terms have the same information content, as for example knowing that an object is a cellular component does not help in categorisation.

Information content of taxonomic terms

The information content of a term tm is given by the negative log-likelihood $ic(tm) = -\log prob(tm)$. The probability of a term $prob(tm)$ can be estimated by the frequency $\widetilde{prob}(tm) = \frac{N(tm)}{N(tm_{root})}$ with $N(tm_{root})$ comprising all terms within a taxonomy. The number $N(tm)$ of a term is hereby the sum of the usage of the term itself $N(tm)$ with the occurrence of its child terms $N(tm) = N(tm) + \sum_{tm' \in \mathbf{N}^{desc}(tm)} N(tm')$. The set $\mathbf{N}^{desc}(tm)$ comprises all first level child nodes of the term tm. The term tm itself is ancestor of these nodes. Given two terms (tm, tm') the closest term which is ancestor to both terms is the closest common ancestor (cca). It defines an upper bound on the similarity measure between these terms.

Methods for calculating the reliability scores between terms

I used four measures that are based on the information content of the cca:

- Resnik [184] proposed to use this measure directly $R^{sem}_{res}(tm, tm') = -\log(\widetilde{prob}(cca))$.
- Lin [130] proposed a normalised measure $R^{sem}_{lin}(tm, tm') = \frac{2 \cdot \log(\widetilde{prob}(cca))}{\log(\widetilde{prob}(tm)) + \log(\widetilde{prob}(tm'))}$.
- Schlicker et al. [187] combines the two previously mentioned methods resulting in a measure of $R^{sem}_{rel}(tm, tm') = \frac{2 \cdot \log(\widetilde{prob}(cca))}{\log(\widetilde{prob}(tm)) + \log(\widetilde{prob}(tm'))} \cdot (1 - \widetilde{prob}(acc))$.
- Jiang and Conrath [101] proposed to measure the similarity as $R^{sem}_{jac}(tm, tm') = 1 - (-\log(\widetilde{prob}(tm)) - \log(\widetilde{prob}(tm'))) + 2 \cdot \log(\widetilde{prob}(cca)$, by considering the link strength between a child and its parent as the difference of their information content.

Furthermore, I used a fifth measure proposed by Wang et al. [225], that also uses the topology of the taxonomy. They define the directed acyclic graph (DAG) of a term tm as $DAG_{tm} = (tm, \mathbf{N}^{anc}_{all}(tm), \mathbf{E}_{tm})$, with $\mathbf{N}^{anc}_{all}(tm)$ being the set of all ancestor terms of tm, including tm. $\mathbf{E}_{tm}$ is the set of edges. A DAG is a directed graph without cycles. The defined DAG comprises the paths from tm to the root term. For each term $tm' \in \mathbf{N}^{anc}_{all}(tm)$ the semantic value $sv_{tm}(tm')$ is calculated as

$$sv_{tm}(tm') = \begin{cases} sv_{tm}(tm) = 1 \\ sv_{tm}(tm') = \max(\zeta_e \cdot sv_{t_1}(tm'') | tm'' \in \mathbf{N}^{anc}_{all}(tm)) \text{ if } tm \neq tm' \end{cases} \tag{21}$$

with $0 < \zeta_e < 1$ as semantic contribution factor, which is predefined based on the kind of relation described by the edge e. The sum of the individual semantic values $osv(tm) = \sum_{tm' \in \mathbf{N}^{anc}_{all}(tm)} sv_{tm}(tm')$ defines the overall semantic value of tm.

The similarity between two terms tm and tm' is defined as

$$R^{sem}_{wan}(tm, tm') = \frac{\sum_{tm'' \in \mathbf{N}^{anc}_{all}(tm) \cap \mathbf{N}^{anc}_{all}(tm')} (sv_{tm}(tm'') + sv_{tm'}(tm''))}{osv(tm) + osv(tm')}. \tag{22}$$

Calculating the reliability scores between genes

If two genes are annotated by several terms the semantic similarity between the genes is estimated using a Best-Match-Average strategy. Let $\mathbf{TM}_1 = \{tm_1^1, tm_1^2, \ldots\}$ and $\mathbf{TM}_2 = \{tm_2^1, tm_2^2, \ldots\}$ being the term sets annotating the genes g_1 and g_2, the similarity between these genes is

$$R^{sem}(g_1, g_2) = \frac{\sum\limits_{1 \leqslant i \leqslant |\mathbf{TM}_1|} R^{sem}(tm_1^i, \mathbf{TM}_2) + \sum\limits_{1 \leqslant j \leqslant |\mathbf{TM}_2|} R^{sem}(tm_2^j, \mathbf{TM}_1)}{|\mathbf{TM}_1| + |\mathbf{TM}_2|}, \text{ with} \tag{23}$$

$$R^{sem}(tm, \mathbf{TM}) = \max_{1 \leqslant l \leqslant |\mathbf{TM}|} R^{sem}(tm, tm^l).$$

The calculation is performed with the R-package *GoSemSim* [243].

2.4.2.3 Composite scoring by STRING

The reliability of interactions can also be estimated taking genomic information into account, like phylogenetic co-occurrence, gene fusion events and genomic neighbourhood. The idea is that genes under the same evolutionary pressure are more likely to be functionally associated. The scoring is based on the probability to find the interaction in the same pathway in Kyoto Encyclopedia of Genes and Genomes (KEGG) [105]. Furthermore, the association between interactions across organisms is included as additional estimation score and text-mining is performed. The text-mining results are scored using the frequency and distribution of gene names. STRING provides a composite score including these measures in addition to experimental results and other databases. The composite score is calculated in a naive Bayesian way leading to an increased score if an interaction is found by several methods [219]. The network extracted from STRING comprises 9106578 interactions and 17440 proteins. For this analysis STRING v. 10.5 [200] and the *stringdb* R-package [67] were used.

2.5 Datasets and processing

Transcriptomic data was used throughout this thesis. It can be considered as giving insight about a cell's stage or phenotype.

2.5.1 High-throughput techniques and data processing

The two main techniques to measure transcriptomic data in a high-throughput way are microarrays and ribonucleic acid (RNA)-sequencing.

Microarray

Microarrays are constructed of short oligonucleotide sequences (probes) bound to a solid chip platform. These probes are complementary to messenger ribonucleic acid (mRNA) transcripts. The transcripts themselves are preprocessed and translated to complementary deoxyribonucleic acid (cDNA). They are labelled with a fluorescent dye and afterwards given on the chip on which they hybridise with their corresponding probes. Often several probes represent the same transcript. The amount of bound transcripts can be measured based on the fluorescence labelling.
As the absolute intensity might depend on the washing or the polymerase chain reaction (PCR) conditions, this method does not quantify transcripts in an absolute way. In order to compare transcript levels and experiments from different laboratories the results are normalised [144].

There are different types and producers of microarray chips. If raw data generated on a chip produced by Affymetrix was provided, I normalised the raw data myself, otherwise if no raw data was provided or the data was generated using another platform I used the normalised data provided in the original publication. The two common normalisation methods for Affymetrix Chips are the mas5 or the log scale robust multi-array average (rma) normalisation. An Affymetrix GeneChip has two kinds of probes, a perfect match probe (PMP), of which the sequence corresponds to the feature, and a mismatch probe (MP), that has the same sequence except of the centre nucleotide and is intended to measure unspecific binding. mas5 uses a log of the difference of these both measures and furthermore their background. Normalisation is calculated per chip. In contrast to that rma normalises over several chips and additionally does not us the mismatch probes [98]. The corresponding normalisation is noted at each dataset.

RNA-sequencing

RNA-sequencing provides direct measurements of transcripts [144]. The data I analysed is either measured by sequencing by ligation (Abi SOLiD) or by sequencing by synthesis (Illumina). Both techniques measure the amount of mRNA in terms of reads, with each read consisting of one measured mRNA sequence. By aligning the reads to a reference genome, a whole genome of the organism investigated, the reads are mapped to their corresponding genes and the amount of mRNA per gene is gained.

There exist several normalisation methods for RNA-sequencing data. Normalisation is required to reduce bias from the sequencing depth or length of mRNA [1]. I used reads per kilo base per million (RPKM) as normalisation, which refers to reads normalised by the total amount of reads and by the gene length, as long genes most likely have more reads than short genes. Given a set of reads $\mathbf{R}$ the RPKM for gene g is calculated as

$$\mathrm{RPKM}(g) = \frac{|\mathbf{R}_g| \cdot 10^6 \cdot 10^3}{|\mathbf{R}| \cdot len_g}, \tag{24}$$

with $\mathbf{R}_g$ being the set of reads mapped to g and len_g being the length of the gene in base pairs.

2.5.2 Gene enrichment analysis

Gene enrichment analysis is performed on a group of selected genes. It reveals whether there is an association between these genes being selected and these genes being annotated by a specific property. I analysed the association with reference to belonging to specific KEGG pathways [105]. The enrichment is calculated using a hypergeometric distribution. It describes the probability that the number of selected genes belonging to one pathway is equal or larger than the measured once drawing as many genes as the pathway has randomly without replacement from all genes under consideration. The Benjamini-Hochberg procedure [26] is used to adjust the p-value and control false discovery rate (FDR). For the calculation the STRING v. 10.5 [200] and the *stringdb* R-package [67] were used.

2.5.3 Dataset characteristics

I analysed ten different datasets. The datasets were either downloaded from Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) [61, 21], ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) [15] or taken from the original publication.

The characteristics of all datasets used in this thesis are given in Table 1.

Table 1: Characteristics of the real datasets. For each dataset the dimensionality, number of features n, the number of classes $|y|$, the overall number of samples m and the number of samples m_i per class y_i is shown. The datasets are ordered and grouped by datasets consisting of static stages, datasets comprising developmental stages and datasets with pre-cancer stages.

n	$\|y\|$	m	m_i					
d_1: **Hypothalamic-pituitary-adrenal (HPA)-axis** [7]								
54	7	336	$m_0 = 48$	$m_1 = 48$	$m_2 = 48$	$m_3 = 48$	$m_4 = 48$	$m_5 = 48$
			$m_6 = 48$					
d_2: **Leukaemia** [111]								
54613	18	2096	$m_0 = 40$	$m_1 = 36$	$m_2 = 58$	$m_3 = 48$	$m_4 = 28$	$m_5 = 351$
			$m_6 = 38$	$m_7 = 37$	$m_8 = 40$	$m_9 = 237$	$m_{10} = 122$	$m_{11} = 448$
			$m_{12} = 76$	$m_{13} = 13$	$m_{14} = 206$	$m_{15} = 74$	$m_{16} = 70$	$m_{17} = 174$
d_3: ***Drosophila melanogaster*** [10]								
4028	4	67	$m_0 = 31$	$m_1 = 10$	$m_2 = 18$	$m_3 = 8$		
d_4: ***Danio rerio*** [210]								
15502	5	69	$m_0 = 14$	$m_1 = 14$	$m_2 = 15$	$m_3 = 12$	$m_4 = 14$	
d_5: ***Caenorhabditis elegans (C. elegans) (stages)*** [24]								
22548	5	62	$m_0 = 12$	$m_1 = 12$	$m_2 = 19$	$m_3 = 7$	$m_4 = 12$	
d_6: ***C. elegans (time points)*** [24]								
22548	10	62	$m_0 = 6$	$m_1 = 6$	$m_2 = 6$	$m_3 = 6$	$m_4 = 6$	$m_5 = 8$
			$m_6 = 5$	$m_7 = 7$	$m_8 = 5$	$m_9 = 7$		
d_7: **Colon carcinoma** [176]								
22215	4	58	$m_0 = 13$	$m_1 = 17$	$m_2 = 17$	$m_3 = 11$		
d_8: **Pancreatic adenocarcinoma, Microarray** [36]								
21521	5	38	$m_0 = 6$	$m_1 = 6$	$m_2 = 8$	$m_3 = 10$	$m_4 = 8$	
d_9: **Hepatocellular carcinoma** [193]								
48803	10	107	$m_0 = 13$	$m_1 = 8$	$m_2 = 12$	$m_3 = 12$	$m_4 = 11$	$m_5 = 11$
			$m_6 = 5$	$m_7 = 9$	$m_8 = 12$	$m_9 = 14$		
d_{10}: **Pancreatic adenocarcinoma, RNA-Sequencing** [16, 17]								
12599	5	32	$m_0 = 3$	$m_1 = 4$	$m_2 = 5$	$m_3 = 9$	$m_4 = 11$	

The datasets can be grouped within three groups. The first group comprises two datasets with static stages representing either different tissue types or different cancer subgroups.

The second group contains datasets of developmental stages of model organisms. Each stage and hence each class represents samples taken at a specific point in time or during a specific time period during development.
The last group is composed of different cancer type datasets with different diagnostic stages. The stages within this group of datasets have been defined based on histological observations and all datasets have the normal and the carcinoma stage in common. The lesions themselves are defined with a different granularity. The most coarse definition is given for colon carcinoma as adenomas either have a low or high dysplasia [28]. Figure 16 shows an approximate overlap of the stages.

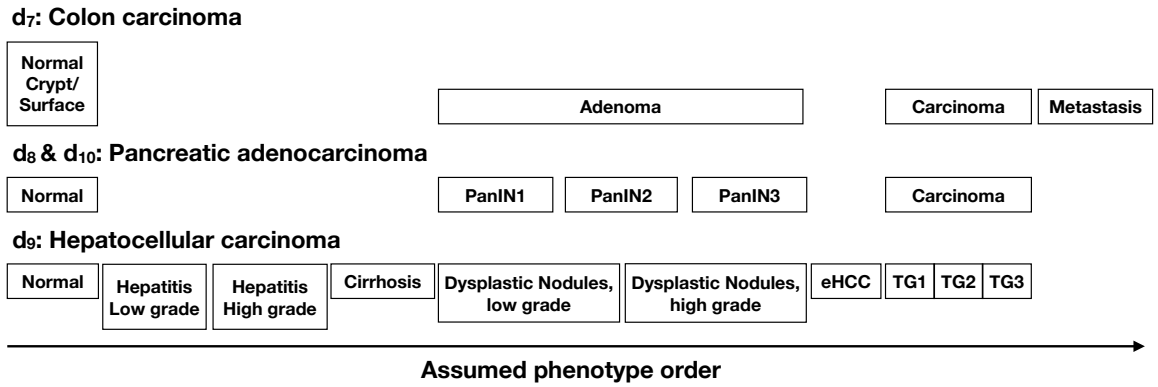


Figure 16: Overlap of pre-cancer stages within three analysed cancer types. The analysed datasets ($d_7 - d_{10}$) comprise three cancer types. The defined and reported precursor stages of the individual cancer types denote different granularities. The stages show a specific overlap. The colon carcinoma dataset for example has one adenoma stage, which is however described by three pancreatic intraepithelial neoplasms (PanINs) in the pancreatic adenocarcinoma (PDAC) and two dysplastic nodule stages in the hepatocellular carcinoma (HCC) dataset. The assumed phenotype order starts with the normal stage and goes towards carcinoma stages via the presumable precursor stages. early HCC (eHCC), tumour grade (TG).

d_1: *HPA-axis*

This dataset comprises seven different classes, corresponding to different tissues of rats. Agba et al. [7] measured DNA methylation patterns of promoters of the glucocorticoid receptor gene (nuclear receptor subfamily 3 group C member 1 (NR3C1)) and the imprinting control region of insulin like growth factor 2/H19 imprinted maternally expressed transcript (IGF2/H19) within these seven tissues using Illumina HiSeq 2500. The preprocessed data can be found on GEO [61, 21] under the accession number GSE103066. The hypothalamic-pituitary-adrenal (HPA)-axis, that is known for stress regulation is enclosed within these data. The classes denote the different tissue types y_0: Cortex, y_1: Hippocampus, y_2: Hypothalamus, y_3: Pituitary,

y_4: Adrenal, y_5: Skin, y_6: Liver.
There is no order on all classes assumed but a subgroup order of Hypothalamus $\prec$ Pituitary $\prec$ Adrenal is assumed.

d_2: *Leukaemia (MILE)*

This dataset was downloaded from GEO [61, 21] under the accession number GSE13159. It was collected within the scope of the MILE Study (Microarray Innovations In LEukaemia) program. Provided by Kohlmann et al. [111] it comprises expression data from blood and bone marrow samples taken from acute and chronic leukaemia patients. The data was measured on the Affymetrix Human Genome U133 Plus 2.0 Array (Affymetrix, Santa Clara, CA, USA), downloaded as raw data and afterwards *mas5* normalised. The samples hereby comprise 18 subgroups labelled as y_0: acute lymphocytic leukaemia (ALL) with hyperdiploid karyotype, y_1: ALL with t(1;19), y_2: ALL with t(12;21), y_3: acute myeloid leukaemia (AML) complex aberrant karyotype, y_4: AML with inv(16)/t(16;16), y_5: AML with normal karyotype & other abnormalities, y_6: AML with t(11q23)/mixed-lineage leukaemia (MLL), y_7: AML with t(15;17), y_8: AMLwith t(8;21), y_9: c-ALL/Pre-B-ALL without t(9;22), y_{10}: c-ALL/Pre-B-ALL with t(9;22), y_{11}: chronic lymphocytic leukaemia (CLL), y_{12}: chronic myeloid leukaemia (CML), y_{13}: mature B-ALL with t(8;14), y_{14}: myelodysplastic syndromes (MDS), y_{15}: Non-leukaemia and healthy bone marrow, y_{16}: Pro-B-ALL with t(11q23)/MLL, y_{17}: T-ALL.
There is no order assumed between the classes.

d_3: *Drosophila melanogaster*

Arbeitman et al. [10] investigated the gene expression profiles of *drosophila melanogaster* by taking samples at different points in time and a cDNA microarray chip constructed using around 9700 *drosophila* cDNA elements. The processed data was available and downloaded from http://flygenome.yale.edu/Lifecycle/. The following classes are included in this dataset: y_0: $\mathtt{embryo}$, y_1: $\mathtt{larva}$, y_2: $\mathtt{pupa}$, y_3: $\mathtt{adult}$.
The assumed order is: $\mathtt{embryo} \prec \mathtt{larva} \prec \mathtt{pupa} \prec \mathtt{adult}$.

d_4: *Danio rerio*

This dataset was downloaded from GEO [61, 21] under the accession number GSE1337 and afterwards rma normalised. Within the study by Toyama et al. [210] gene expression profiles of the pineal glands of *danio rerio* were investigated using the Affymetrix Zebrafish Genome Array (Affymetrix, Santa Clara, CA, USA). Hereby, the points in time collected 3, 5 and 10 days after birth as well as at an age of 3 months and 1-2 years were categories within the following five classes: y_0:

$\mathtt{embryo}_1$, y_1: $\mathtt{embryo}_2$, y_2: $\mathtt{embryo}_3$, y_3: $\mathtt{adult}_1$, y_4: $\mathtt{adult}_2$.
The assumed order is: $\mathtt{embryo}_1 \prec \mathtt{embryo}_2 \prec \mathtt{embryo}_3 \prec \mathtt{adult}_1 \prec \mathtt{adult}_2$.

d_5 *and* d_6*: Caenorhabditis elegans*

The dataset collected by Baugh et al. [24] comprises samples of wild-type and mutant embryos of *Caenorhabditis elegans* (*C. elegans*) taken at 10 points in time after the 4-cell-stage of the embryo. It was measured on Affymetrix *C. elegans* Genome Array (Affymetrix, Santa Clara, CA, USA), downloaded from GEO [61, 21] under the accession number GSE2180 and afterwards rma normalised. The data of the mutant variant without C blastomere was used and once grouped according to the developmental stages y_0: $\mathtt{stage}_1$ (0 and 23 minutes), y_1: $\mathtt{stage}_2$ (41 and 53 minutes), y_2: $\mathtt{stage}_3$ (66, 83 and 101 minutes), y_3: $\mathtt{stage}_4$ (122 minutes), y_4: $\mathtt{stage}_5$ (143 and 186 minutes).
The assumed order is: $\mathtt{stage}_1 \prec \mathtt{stage}_2 \prec \mathtt{stage}_3 \prec \mathtt{stage}_4 \prec \mathtt{stage}_5$.

Additionally, the class labels were assigned based on the points in time taken: y_0: t_1 (0 minutes), y_1: t_2 (23 minutes), y_2: t_3 (41 minutes), y_3: t_4 (53 minutes), y_4: t_5 (66 minutes). y_5: t_6 (83 minutes), y_6: t_7 (101 minutes), y_7: t_8 (122 minutes), y_8: t_9 (143 minutes), y_9: t_{10} (186 minutes).
The assumed order is: $t_1 \prec t_2 \prec t_3 \prec t_4 \prec t_5 \prec t_6 \prec t_7 \prec t_8 \prec t_9 \prec t_{10}$.

d_7*: Colon carcinoma*

Qu et al. [176] investigated the development of colorectal cancer. They performed microarray analysis on laser-captured frozen tissue using the Affymetrix HG-U133A GeneChips (Affymetrix, Santa Clara, CA, USA). The data was downloaded from GEO [61, 21] under the accession number GSE77953. The staging and hence the labelling was performed by a pathologist. The data was rma normalised after the download. y_0: normal crypt and surface epithelium, y_1: adenoma, y_2: carcinoma, y_3: metastasis.
The assumed order is: normal crypt & surface epithelium $\prec$ adenoma $\prec$ carcinoma $\prec$ metastasis.

d_8*: Pancreatic adenocarcinoma, Microarray*

This dataset was gathered and block normalised by Buchholz et al. [36]. It comprises microarray expression measurements of manually microdissected tissue extracted from the head of the pancreas. The used chip comprises oligonucleotides of the Human Genome Oligo-Set Version 2.0 (Operon, Germany) spotted on GAPSII Slides (Corning, USA). The labels are assigned based on histology and according to the WHO classification scheme [92]. For the entropy calculation the expression data was shifted by 0.1 in order to eliminate the zero expression entries, which

might lead to varying sizes of the strongly connected component over the different classes. y_0: normal duct, y_1: PanIN1, y_2: PanIN2, y_3: PanIN3, y_4: pancreatic adenocarcinoma (PDAC).
The assumed order is: normal $\prec$ PanIN1 $\prec$ PanIN2 $\prec$ PanIN3 $\prec$ PDAC.

d_9: *Hepatocellular carcinoma*

This data was retrieved from patients with liver disease within a study of drivers for HCC by Shen et al. [193] using HumanHT-12 V4 Expression BeadChips (Illumina,San Diego, CA) and downloaded quantile normalised with GenPlex 3.0. The data was downloaded from GEO [61, 21] under the accession number GSE89377. The following categories are distinguished: y_0: normal, y_1: Chronic Hepatitis with low grade, y_2: Chronic Hepatitis with high grade, y_3: Cirrhosis, y_4: Dysplastic nodules with low grade, y_5: Dysplastic nodules with high grade, y_6: early HCC, y_7: HCC (tumour grade (TG)1), y_8: HCC (TG2), y_9: HCC (TG3).
The assumed order is: normal $\prec$ Chronic Hepatitis with low grade $\prec$ Chronic Hepatitis with high grade $\prec$ Cirrhosis $\prec$ Dysplastic nodules with low grade $\prec$ Dysplastic nodules with high grade $\prec$ early HCC $\prec$ HCC (TG1) $\prec$ HCC (TG2) $\prec$ HCC (TG3).

d_{10}: *Pancreatic adenocarcinoma, RNA-Sequencing*

This dataset is a combined dataset of two studies investigating gene expression data of laser microdissected pancreatic (pre-)cancer tissue. Ayars et al. [17] measured expression data from laser microdissected tissue of normal duct, PanIN1, PanIN2 and PanIN3. The PDAC samples are also microdissected and part of the COMPASS trial [16]. It is available in the ArrayExpress database under accession number EGAD00001003584 (http://www.ebi.ac.uk/arrayexpress). Access was granted by the Ontario Institute for Cancer Research. The authors provided aligned data. This was further processed using htseq count [9] and GrCh38 [188]. Finally, RPKM values were calculated. The raw data provided by Ayars et al. [17] was aligned using the star aligner [59], similar to the Comprehensive Molecular Characterisation of Advanced Pancreatic Ductal Adenocarcinomas for Better Treatment Selection (COMPASS) trial. Read counting and RPKM calculation was afterwards performed in the same way. Alignment and normalisation was performed by Dr. Axel Fürstberger, who kindly provided the processed data. The classes are: y_0: normal duct, y_1: PanIN1, y_2: PanIN2, y_3: PanIN3, y_4: PDAC.
The assumed order is: normal $\prec$ PanIN1 $\prec$ PanIN2 $\prec$ PanIN3 $\prec$ PDAC.

3

ORDINAL CLASSIFIER CASCADES & PHENOTYPE NETWORK RECONSTRUCTION

3.1 Ordinal classifier cascade - advancements

Ordinal classifiers are usually designed for using one known order on the labels $\mathcal{Y}$ to improve classification performance [173, 69, 38, 68]. Screening for potential orders requires to evaluate not only the performance of one order, but of all possible orderings. As consequence $|\mathcal{Y}|!$ permutations have to be evaluated in the case of a screening for total orders and $|\mathcal{Y}|$ classes, with an even enlarging search space for the inclusion of orders in subsets of classes.
If subsets of classes of size k with $k \in \{1, \ldots, |\mathcal{Y}|\}$ are considered, the search space for $|\mathcal{Y}|$ classes increases from $f_o^{total}(|\mathcal{Y}|) = |\mathcal{Y}|!$ to $f_o^{all}(|\mathcal{Y}|) = \sum_{k=1}^{|\mathcal{Y}|} \binom{|\mathcal{Y}|}{k} k!$. This results in the number of subcascades of size k as being $f_o^k(|\mathcal{Y}|) = \binom{|\mathcal{Y}|}{k} k!$. Subcascades of length $k = 1$ are included for simplicity reasons.
Interestingly, the ratio of increase between the number of full and subcascades can be approximated by the Euler number $\frac{f_o^{all}(|\mathcal{Y}|)}{f_o^{total}(|\mathcal{Y}|)} = \sum_{k=1}^{|\mathcal{Y}|} \frac{1}{k!} \xrightarrow{|\mathcal{Y}| \to \infty} e \sim 2.718$, as

$$\begin{aligned} f_o^{all}(|\mathcal{Y}|) &= \sum_{k=1}^{|\mathcal{Y}|} \binom{|\mathcal{Y}|}{k} k! = \sum_{k=1}^{|\mathcal{Y}|} \frac{|\mathcal{Y}|!}{k!(|\mathcal{Y}| - k)!} k! = |\mathcal{Y}|! \sum_{k=1}^{|\mathcal{Y}|} \frac{1}{(|\mathcal{Y}| - k)!} \\ &= |\mathcal{Y}|! \sum_{k=1}^{|\mathcal{Y}|} \frac{1}{k!} \end{aligned} \tag{25}$$

For a dataset comprising 5 classes the evaluation of all cascades and subcascades sums up to 300. Increasing the number of classes to 10 almost 10 million evaluations have to be performed, showing the need of a fast and efficient screening procedure.

In the following I will introduce two strategies to deal with the large search space and the high computational demand: On the one hand a preselection of potential candidate cascades leading to a search space reduction, and on the other hand a precalculation scheme.

3.1.1 Preselection strategy

In order to screen for a reflected order, all permutations of all classes have to be analysed for their performance. Theorem 1 shows that the overall performance in terms of minimal class-wise sensitivity is upper bounded by the performance of the ordinal classifier cascade (Occ)'s base classifier. It is a consequence of Appendix Lemmata A1 and A2. A consequence of this theorem is, that many permutations can be sorted out by the use of a threshold on the base classifiers' performances. This threshold acts as minimal requirement on the overall performance [123] .

Theorem 1. [123] Let h denote an Occ $h : \mathbb{R}^n \longrightarrow \mathcal{Y} = \left\{y_{(1)}, \ldots, y_{(|\mathcal{Y}|)}\right\}$ with base classifiers $\mathcal{E} = \left\{c_{(1)}, \ldots, c_{(|\mathcal{Y}|-1)}\right\}$. Let furthermore $\mathbf{X}_{y_{(i)}}$ be a non-empty set of samples of class $y_{(i)}$. Then the sensitivity of h for $y_{(i)}$ is limited by

$$p_h(y_{(i)} \,|\, \mathbf{X}_{y_{(i)}}) \leqslant p_{c_{(i)}}(y_{(i)} \,|\, \mathbf{X}_{y_{(i)}}) \tag{26}$$

$$p_h(y_{(i)} \,|\, \mathbf{X}_{y_{(i)}}) \leqslant \min_{k<i} p_{c_{(k)}}(y_{(k+1)} \,|\, \mathbf{X}_{y_{(i)}}). \tag{27}$$

As mentioned in the previous section the search space for an ordinal classifier consists of all possible permutations. However, by the use of this theorem not all permutations have to be included into the final evaluation. Instead possible permutations can be built up incrementally and sorted out by the use of requirements on the base classifiers' performances. Those requirements can be ascribed to Theorem 1 and are a key aspect of the CASCADES algorithm (Algorithm 1) established, enhanced and evaluated in this thesis and published in Lausser* et al. [123].

Starting with an empty set of candidate cascades $\mathcal{C} = \emptyset$ and the complete set of labels $\mathcal{Y}$, each label y is a potential start of one cascade. Building up each possible cascade incrementally y_i will be denoted as the previous class label, y_j as the potential next class label and t as the threshold on the minimal class-wise sensitivity. For appending a possible new label y_j, the performance of the base classifier c_{ij} is evaluated. The label y_j is only appended to the previous cascade (preamble) if the requirements on the sensitivity

$$FC_{ij} = p_{c_{ij}}(y_i \,|\, \mathbf{X}_i) \geqslant t \tag{28}$$

of the current class (y_i) and the conditional prediction rates

$$\forall\, y_l \in \mathcal{Y} : SC_{ij}(l) = p_{c_{ij}}(y_j \,|\, \mathbf{X}_l) \geqslant t \tag{29}$$

of the remaining classes ($y_r \in \mathcal{Y} \setminus y_i$) are fulfilled.

Algorithm 1 CASCADES($\mathcal{Y}$, $\mathcal{C}$, y_i, t). The parameters used are $\mathcal{Y}$ for the set of class labels, $\mathcal{C}$ as set of candidate cascades, y_i depicting the current class label and t the threshold describing the requirement on the minimal sensitivity. Furthermore $FC_{i,j}$ is the first class sensitivity, hence the sensitivity of class y_i using the base classifier c_{ij} and $SC_{i,j}(l)$ is the percentage of samples of class y_r being predicted as y_j using c_{ij}.

```
1: if |𝒴| = 0 then
2:    return 𝒞
3: else
4:    𝒞_s := ∅
5:    if y_i = ∅ then
6:       for y_j ∈ 𝒴 do
7:          𝒞_c := CASCADES(𝒴 \ {y_j}, {(y_j)}, y_j, t)
8:          𝒞_s := 𝒞_s ∪ 𝒞_c
9:       end for
10:      return C_s
11:   else
12:      for y_j ∈ 𝒴 do
13:         if FC_{i,j} ⩾ t and ∀ y_r ∈ 𝒴 : SC_{i,j}(l) ⩾ t then
14:            𝒞_r := {(y_(1), ..., y_(k), y_j) | (y_(1), ..., y_(k)) ∈ 𝒞}
15:            𝒞_c := CASCADES(𝒴 \ {y_j}, 𝒞_r, y_j, t)
16:         else
17:            𝒞_c := ∅
18:         end if
19:         𝒞_s := 𝒞_s ∪ 𝒞_c
20:      end for
21:      return C_s
22:   end if
23: end if
```

The condition $FC_{ij} \geqslant t$ and $\forall y_l \in \mathcal{Y} : SC_{ij}(l) \geqslant t$ control whether the sensitivity of the current class y_i passes the threshold and whether the current classifier lets enough samples of the remaining classes pass by predicting them as y_j. The sensitivities and conditional prediction rates only depend on the pairwise training and can be saved in a look-up table (Table 2).
If the requirements are not fulfilled the whole preamble cascade can be removed from the set of cascades. Also this class can be removed from $|\mathcal{Y}|$ and the next label can be checked. In the end a set of possible cascades is returned.

Table 2: Look-up table for the preselection of possible orderings. Given a dataset comprising $\mathcal{Y}$ classes. For each binary base classifier $c_{ij}(y_i, y_j \in \mathcal{Y}, i \neq j)$ trained on samples $\mathbf{x} \in \mathbf{X}_i \cup \mathbf{X}_j$, with $\mathbf{X}_i$ comprising all samples of class y_i and $\mathbf{X}_j$ all samples of class y_j, respectively, the first class sensitivity $FC_{ij} = p_{c_{ij}}(y_i \mid \mathbf{X}_i)$ and all conditional prediction rates $SC_{ij}(l) = p_{c_{ij}}(y_j \mid \mathbf{X}_l) \forall l \in \mathcal{Y}$ are given. See [122] licensed under CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

$c_{ij}(\mathbf{x})$ trained on class y_i, y_j		first class FC_{ij}	second class $SC_{ij}(l)$				
		y_i	y_1	$\dots$	y_l	$\dots$	$y_{\lvert\mathcal{Y}\rvert}$
y_1	y_2	$p_{c_{12}}(y_1 \mid \mathbf{X}_1)$	$p_{c_{12}}(y_2 \mid \mathbf{X}_1)$	$\dots$	$p_{c_{12}}(y_2 \mid \mathbf{X}_l)$	$\dots$	$p_{c_{12}}(y_2 \mid \mathbf{X}_{\lvert\mathcal{Y}\rvert})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
y_i	y_j	$p_{c_{ij}}(y_i \mid \mathbf{X}_i)$	$p_{c_{ij}}(y_j \mid \mathbf{X}_1)$	$\dots$	$p_{c_{ij}}(y_j \mid \mathbf{X}_l)$	$\dots$	$p_{c_{ij}}(y_j \mid \mathbf{X}_{\lvert\mathcal{Y}\rvert})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$

As this is however only a preselection based on an upper bound and hence the final cascades might show a lower performance this remaining set has to be evaluated using the Occ to result in a final performance measure.

Due to the preselection not all permutations have to be screened explicitly, which shortens screening time and allows for screening datasets with high number of classes. The final workflow for using the Occ can be divided into first the pairwise training, the thinning out of the search space or the built up of potential orders and the evaluation of these orders. An overall scheme is given in Figure 17.

3.1.2 Look-up structure - prediction map

Computational complexity can further be reduced by the newly introduced look-up structure [122]. Instead of only storing the sensitivities and conditional prediction rates (Table 2), the predictions of each binary base classifier c_{ij} for each validation sample $(\mathbf{x}, y) \in \mathbf{V}$ can be memoized in a prediction map (Table 3). The sensitivities and conditional prediction rates used for the preselection of potential cascades can then be derived from this prediction map. The advantages of this map are that firstly these predictions can also be used for the final evaluation of the remaining cascades and hence no de-novo training of the base classifiers is needed and secondly once the prediction map for all classes $|\mathcal{Y}|$ is calculated it can be easily reduced to subsets of $\mathcal{Y}$.

Table 3: Prediction map memoizing the predictions of all binary classifiers c_{ij}. For each binary base classifier c_{ij}, with $y_i, y_j \in \mathcal{Y}, i \neq j$ its predictions on all samples $(\mathbf{x}, y)$ in the validation set $\mathbf{V}$ are denoted. See [122] licensed under CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

binary	validation samples $(\mathbf{x}, y) \in \mathbf{V}$				
classifier	$\mathbf{x} \in \mathbf{X}_{y_1}$	...	$\mathbf{x} \in \mathbf{X}_{y_k}$	...	$\mathbf{x} \in \mathbf{X}_{y_{\lvert\mathcal{Y}\rvert}}$
$c_{ij}(\mathbf{x})$	$\mathbf{x}_1 \dots \mathbf{x}_{\lvert\mathbf{X}_{y_1}\rvert}$	...	$\mathbf{x}_1 \dots \mathbf{x}_{\lvert\mathbf{X}_{y_k}\rvert}$	...	$\mathbf{x}_1 \dots \mathbf{x}_{\lvert\mathbf{X}_{\lvert\mathcal{Y}\rvert}\rvert}$
y_1, y_2	$c_{12}(\mathbf{x}_1) \dots c_{12}(\mathbf{x}_{\lvert\mathbf{X}_{y_1}\rvert})$	...	$c_{12}(\mathbf{x}_1) \dots c_{12}(\mathbf{x}_{\lvert\mathbf{X}_{y_k}\rvert})$	...	$c_{12}(\mathbf{x}_1) \dots c_{12}(\mathbf{x}_{\lvert\mathbf{X}_{y_k}\rvert})$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
y_i, y_j	$c_{ij}(\mathbf{x}_1) \dots c_{ij}(\mathbf{x}_{\lvert\mathbf{X}_{y_k}\rvert})$	...	$c_{ij}(\mathbf{x}_1) \dots c_{ij}(\mathbf{x}_{\lvert\mathbf{X}_{y_k}\rvert})$	...	$c_{ij}(\mathbf{x}_1) \dots c_{ij}(\mathbf{x}_{\lvert\mathbf{X}_{y_k}\rvert})$
$\vdots$	$\vdots$		$\vdots$		$\vdots$

3.1.3 Performance relations between cascades and their subcascades

Analysing subcascades of different lengths results in two objectives, the subcascade length and the performance. Longer cascades constrain the feature space more and might be assumed as more informative, but at the same time reaching a high performance is easier for short subcascades, as they are lower bounded by the performance of their elongations. This bound is given by Theorem 2 published in Lausser* et al. [122]. The proof is additionally given in the Appendix Proof 7.

Theorem 2. [122] Let $y_{(1)} \prec \dots \prec y_{(i)} \prec \dots \prec y_{(|\mathcal{Y}|)}$ denote an arbitrary but fixed class order of the classes in $\mathcal{Y}$ and $h : \mathbb{R}^n \to \mathcal{Y}$ an Occ designed for this class order and based on an ensemble of pairwise trained base classifiers $\mathcal{E} = \{c_{(i)}\}_{i=1}^{|\mathcal{Y}|-1}$. Let furthermore $\mathcal{Y}' = \mathcal{Y} \setminus \{y_{(i)}\}$ and $h' : \mathbb{R}^n \to \mathcal{Y}'$ denote an Occ designed for class order $y_{(1)} \prec \dots \prec y_{(i-1)} \prec y_{(i+1)} \prec \dots \prec y_{(|\mathcal{Y}|)}$ based on $\mathcal{E}' = \mathcal{E} \setminus \{c_{(1)}\}$, if $i = 1$, $\mathcal{E}' = \mathcal{E} \setminus \{c_{(|\mathcal{Y}|-1)}\}$, if $i = |\mathcal{Y}|$ and

$$\mathcal{E}' = \left(\mathcal{E} \setminus \{c_{(i-1)}, c_{(i)}\}\right) \cup \{c'_{(i-1)}\} \tag{30}$$

otherwise, where

$$c'_{(i-1)}(\mathbf{x}) := \begin{cases} y_{(i-1)} & \text{if} \quad c_{(i-1)}(\mathbf{x}) = y_{(i-1)} \\ y_{(i+1)} & \text{if} \quad c_{(i-1)}(\mathbf{x}) = y_{(i)} \end{cases}. \tag{31}$$

In this case

$$\mathrm{smin}_h(\mathcal{Y}) \leqslant \mathrm{smin}_{h'}(\mathcal{Y}'). \tag{32}$$

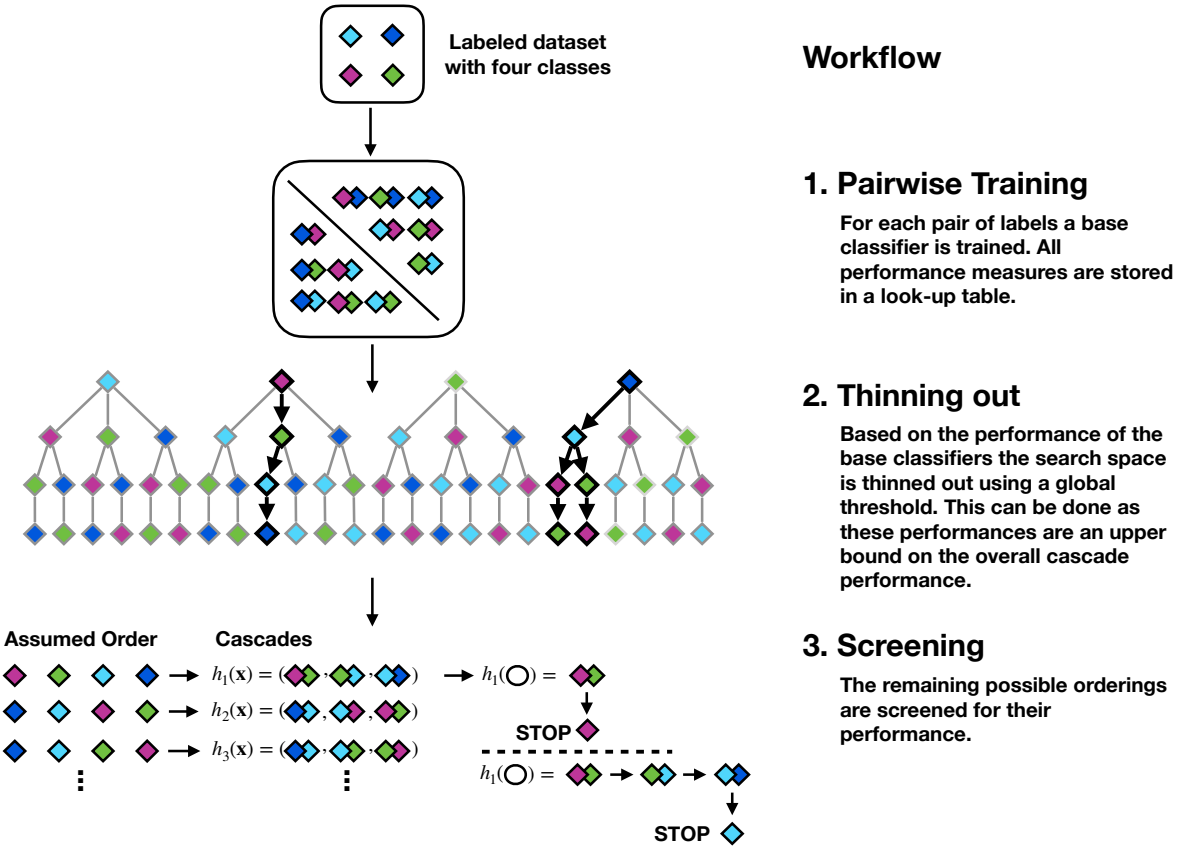


Figure 17: Screening workflow of the ordinal classifier cascade (Occ). Depicted is a four class example. Each class is represented by a coloured square. The first step is a pairwise training on each class combination. Based on the performance of these base classifiers the search space can be thinned out based on the upper bounds shown in Theorem 1. Only the remaining orders have to be evaluated using the Occ. See [123] licensed under CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

3.1.4 R-package *ORION*

The screening functions needed to use the Occs have been bundled into an R-package, called *ORION* [121], that is available from Comprehensive R Archive Network (CRAN). The core functions of this package are the calculation of a prediction map, which is the look-up table given in Table 3 and the screening for (sub)cascades using the CASCADES-algorithm. Furthermore, it contains many functions for analysing and filtering the set of returned cascades, like imposing further thresholds on the minimal class-wise sensitivity or searching for all cascades including a given ordinal pattern.
In general, the package allows for easily screening, filtering, analysing and visualising possible orderings and their characteristics within diverse datasets.

3.2 Ordinal classifier cascade - dependency on dimensionality

As not only the type of base classifier might influence the returned set of cascades but also the dimensionality of the data, I analysed the influence of the dimensionality on the Occ. Based on Cover's theorem [51] it is known that the probability of finding a perfect linear separation of two classes for any random labelling of samples in general position increases with the dimensionality. He showed that the separation capacity of surfaces with n degrees of freedom equals $2n$. Furthermore, the probability that m points in general position with two random assigned labels are separable in $n = \frac{m+\nu\sqrt{m}}{2}$ dimensions follows a cumulative normal distribution

$$\Phi(\nu) = \frac{1}{\sqrt{2\pi}} \int_{-\inf}^{\nu} \exp^{\frac{-x^2}{2}} dx \quad \text{with} \quad \nu = (n - \frac{m}{2}) \cdot \frac{2}{\sqrt{m}}. \tag{33}$$

This illustrates the curse of dimensionality: if the number of dimensions is high compared to the sample number the probability of finding a separation function for any kind of labelling is high.

By performing classification experiments on randomly labelled datasets with increasing dimensionality I was able show that in contrast to the *one-against-one* (OaO) fusion approach the Occ is not sensitive to the dimensionality. By a decomposition it could be further shown that the external rates used for testing but not for training are responsible for this prevention.

3.2.1 Experimental results

I sampled three datasets comprising varying number of classes $|y| \in \{5, 7, 10\}$. The sample set of each class consisted of m labelled samples $\mathbf{X}_y = \{(\mathbf{x}_i, y)\}_{i=1}^{m}$ which were drawn according to a n-dimensional uniform distribution $\mathbf{x} \sim \mathcal{U}(0,1)^n$. I constructed each dataset 1000 times for each dimensionality $n \in [10; 10^4]$.

The results of the reclassification experiments with the OaO approach and the Occ are shown in Figure 18. The fraction of perfect separation in case of the OaO classifier system relates to separations showing a minimal class-wise sensitivity of all classes of 1 and in case of the Occ to cascades showing a minimal class-wise sensitivity of 1. Additionally, the performance of the underlying base classifiers were reported and reproduced the results $\Phi(\alpha)$ shown by Cover [51]. It can be seen that the OaO approach follows

$$\tilde{p}(y) = \Phi(\nu)^{\frac{|y|(|y|-1)}{2}}, \tag{34}$$

with an exponent corresponding to the number of dichotomies within the multiclass experiment.

With increasing number of binary separations resulting in a perfect result, also the OaO approach could separate perfectly. In contrast to this the Occ did not return any cascade even for high dimensionality.

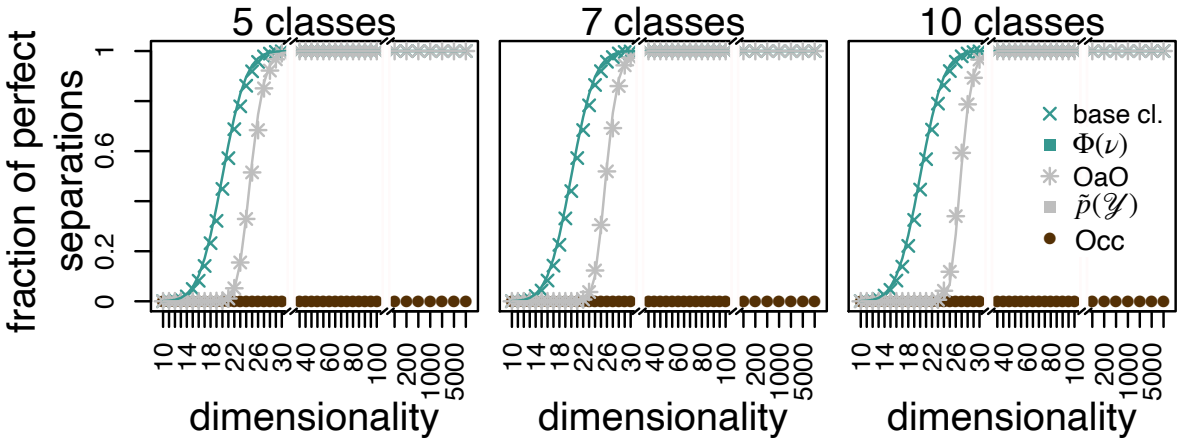


Figure 18: Classification performance dependency by increasing dimensionality for the ordinal classifier cascade (Occ) and the *one-against-one* (OaO) multi-class fusion architecture. Denoted is the fraction of perfect separation counted as a minimal class-wise sensitivity of 1 in randomly sampled artificial datasets comprising 5, 7 and 10 classes. The measured binary base classifier (base cl.) performances follow the curve $\Phi(\nu)$ shown by Cover [51], the measured results of the OaO approach follow $\tilde{p}(\mathcal{Y})$ (Equation 34).

3.2.2 Theoretical results

The observed results could theoretically be shown by analysing the conditional probabilities $p_c(y|y')$ of predicting class label y if the sample was taken from class y' for all binary classifiers c. The class-wise sensitivity for the Occ is a product of class sensitivities (cl.sens.) and external rates (ext.rates)

$$p_{h_{Occ}}(y_{(i)}|y_{(i)}) = p_{c_{(i)}}(y_{(i)}|y_{(i)}) \prod_{l<i} p_{c_{(l)}}(y_{(l+1)}|y_{(i)}) \tag{35}$$

$$= \underbrace{p_{c_{(i)}}(y_{(i)}|y_{(i)})}_{\text{cl.sens.}} \underbrace{p_{c_{(i-1)}}(y_{(i)}|y_{(i)})}_{\text{cl.sens.}} \prod_{l<i-1} \underbrace{p_{c_{(l)}}(y_{(l+1)}|y_{(i)})}_{\text{ext.rates}} \tag{36}$$

For technical reasons it is defined that $p_{c_{(i)}} = 1$ if $c_{(i)} \notin \mathcal{E}_{Occ}$, with $\mathcal{E}_{Occ} = \{c_{(i)}\}_{i=1}^{|\mathcal{Y}|-1}$.

Within an Occ two base classifiers $c_{(i)}$ and $c_{(i-1)}$ are trained on predicting class label $y_{(i)}$. As the base classifiers are evaluated sequentially predicting class label

$y_{(i)}$ also depends on the external rates of all base classifiers $c_{(l<i-1)}$, in order to reach the two base classifiers $c_{(i)}$ and $c_{(i-1)}$ in the sequence.

The class-wise sensitivity of an OaO classifier approach depends on the sensitivities of the base classifiers that can predict y (cl.sens) and on the combination of predictions from the remaining set of base classifiers (ext.rates). It hence can be decomposed as

$$p_{h_{OaO}}(y|y) = \sum_{l=\lceil \frac{|\mathcal{O}|}{2} \rceil}^{|\mathcal{O}|} \sum_{\substack{\mathcal{Y}_1 \subseteq \mathcal{O} \\ |\mathcal{Y}_1|=l}} \left(\prod_{y' \in \mathcal{Y}_1} \underbrace{p_{c_{yy'}}(y|y)}_{\text{cl.sens.}} \right) \left(\prod_{\substack{\mathcal{Y}_0 = \mathcal{O} \setminus \mathcal{Y}_1 \\ y'' \in \mathcal{Y}_0}} 1 - \underbrace{p_{c_{yy''}}(y|y)}_{\text{cl.sens.}} \right) \underbrace{p_{\text{rest}}(y, \mathcal{Y}_1, \mathcal{Y}_0)}_{\substack{\text{ext.rates} \\ \leqslant 1}}. \tag{37}$$

Only base classifiers that are trained on y can predict y. This set of classifiers comprises pairwise classifiers trained on y against all remaining classes $\mathcal{O} = \mathcal{Y} \setminus \{y\}$. Assuming $y' \in \mathcal{O}$, a base classifier within this set either predicts y correctly if $y' \in \mathcal{Y}_1$ or not if $y' \in \mathcal{Y}_0$. The correct prediction of y furthermore depends on the prediction pattern of the classifiers that cannot predict y, which is only based on external rates. This is taken into account by $p_{\text{rest}}(y, \mathcal{Y}_1, \mathcal{Y}_0)$, which can be interpreted as probability that this pattern does not define the overall predicted classes. At least $\lceil \frac{|\mathcal{O}|}{2} \rceil$ base classifiers have to predict y correctly, otherwise the probability of correct prediction is 0.

Using this decomposition it could be shown that $p_{h_{OaO}}(y|y)$ is bounded by the sensitivities of its base classifiers and is hence up to a certain amount unaffected by external rates (Appendix Proof 7).

The bounds of $p_{h_{OaO}}(y|y)$ are

$$\sum_{i=\lceil \frac{|\mathcal{O}|}{2} \rceil}^{|\mathcal{O}|} \binom{|\mathcal{O}|}{i} p_{max}^{i} (1 - p_{min})^{|\mathcal{O}|-i} \geqslant p_{h_{OaO}}(y|y) \geqslant p_{min}^{|\mathcal{O}|}, \tag{38}$$

with p_{max} and p_{min} being the highest and smallest sensitivity of all base classifiers that can predict y, respectively. The lower bound is defined based on the assumption that all classifiers that can predict y, predict y correctly ($\mathcal{Y}_1 = \mathcal{O}$). In this case the patterns resulting from the classifiers that cannot predict y cannot overrule the prediction anymore. The overall sensitivity relies only on the first term comprising the product of class sensitivities and is hence lower bounded by the smallest sensitivity.

The upper bound is also solely based on sensitivities as $p_{rest}(y, \mathcal{Y}_1, \mathcal{Y}_0) \leqslant 1$ and hence relies only on the performance of those base classifiers that can predict y. The binomial coefficient $\binom{|\mathcal{O}|}{i}$ refers to the number of class combination possibilities with $|\mathcal{Y}_1| = i$. The maximal and minimal sensitivities (p_{max}, p_{min}) can be used for overestimating and underestimating the single prediction rates, respectively.

3.3 Ordinal classifier cascade - characterisation

To better understand the functionality of the Occ and the relation between returned cascades and class relations in feature space, I sampled artificial data that represent gene expression data of related classes with various types of relational patterns. This data is supposed to represent processes in which any phenotype can have only one predecessor but several successors in order to fit to the biological situation of cell or tissue-stage relations during development or cancer progression.

3.3.1 Simulating artificial data

Artificial data was based on a set of classes $\mathcal{Y} = \{y_1, ..., y_{|Y|}\}$ and their underlying (partial) ordinal relations. These relations could be expressed as a set of pairwise dependencies $\mathbf{O} = \{(y \prec y')_i | y \neq y'\}_{i=1}^{o}$, with o being the smallest number of pairwise relations needed to describe the given (partial) ordinal relation unambiguously. Within a relation $y \prec y'$ I will call y predecessor and y' successor.
Each of the $|\mathcal{Y}|$ classes comprised m samples. Samples represent the expression values on a log-scale.

3.3.1.1 Prototype generation

Given a set of classes $\mathcal{Y}$ for each class a prototype was sampled resulting in a set of prototypes $\mathbf{P}_{\mathcal{Y}} = \{\mathbf{p}_{y_l}\}_{l=1}^{|\mathcal{Y}|}$, $\mathbf{p}_{y_l} \in \mathbb{R}^n$.
If no predecessor was defined for a class y_l its prototype $\mathbf{p}_{y_l} = (p_{y_l}^1, \ldots, p_{y_l}^n)$ was drawn from a multinomial normal distribution

$$
\begin{aligned}
&\mathbf{p}_{y_l} = (p_{y_l}^1, \ldots, p_{y_l}^n), \text{with} \\
&p_{y_l}^i \sim \mathcal{N}(\mu_p^i, \sigma_p^i), \text{ with } \mu_p^i, \sigma_p^i \in \mathbb{R} \text{ and} \\
&\mu_p^i \sim \mathcal{N}(\mu_\mu, \sigma_\mu), \text{ with } \mu_\mu, \sigma_\mu \in \mathbb{R}, \\
&\sigma_p^i \sim \Gamma(\mu_\sigma, \sigma_\sigma), \text{ with } \mu_\sigma, \sigma_\sigma \in \mathbb{R}.
\end{aligned}
\tag{39}
$$

The prototypes for the remaining classes were sampled in dependency on the prototypes of their direct predecessors. Given the relation $y_i \prec y_j$ and the prototype $\mathbf{p}_{y_i}$ the prototype $\mathbf{p}_{y_j}$ was assumed to differ by numerous features/genes being

differentially expressed. The distribution of the log2 fold change could be assumed to follow a gamma distribution $\Gamma(\alpha, \beta)$, with α and β being the shape and rate parameter, respectively [251]. It followed that

$$\begin{aligned} &\mathbf{p}_{y_j} = \mathbf{p}_{y_i} + \mathbf{r}_{y_j} \circ \mathbf{f}_{y_j}, \text{ with} \\ &\mathbf{f}_{y_j} = (f^1_{y_j}, \ldots, f^n_{y_j}), \mathbf{f}_{y_j} \in \mathbb{R}^n \text{ and } f^l_{y_j} \sim \Gamma(\alpha, \beta) \end{aligned} \tag{40}$$

and $\circ$ denoting the Hadamard-product, and the regulation parameter defining the direction of the change $\mathbf{r}_{y_j} \in \{-1, 1\}^n$.

3.3.1.2 Sample generation

For each class y a set of samples $\mathbf{T}_y$ is generated assuming a multivariate normal distribution

$$\mathbf{T}_y = \{(\mathbf{x_i}, y) | \mathbf{x_i} \sim \mathcal{N}(\mathbf{p_y}, \Sigma_y)\}_{i=1}^m. \tag{41}$$

Genes interact with each other and all interactions span a network, with genes as nodes and their relations as edges. This network can be described by its adjacency matrix G, with entry e_{ij} denoting the interaction between gene i and j. Under the assumption that the covariance matrix gives insights about the relations between the features/genes and considering that the topology of biological networks is described as scale-free, I based the covariance matrix on an adjacency matrix of a scale-free graph. I used the Barabasi-Albert model, which was shown to generate scale-free models, to sample the adjacency matrix. A scale-free network is characterised by its degree distribution $de(deg)$ following a power low $de(deg) \sim deg^{-\gamma}$, with deg being the degree and describing the number of connections per node and $2 < \gamma < 3$ [20].

To describe positive and negative relations the edge weights were randomly assigned as -1 or 1. The same graph was used for all classes and its adjacency matrix is denoted as A. To account for variances in the calculation the main diagonal was set to 1.

In analogy to other artificial data [115], I sampled the feature variance as $\boldsymbol{\sigma}^2_y = (\frac{\sigma^{2^1}_y}{(m-1)}, \ldots, \frac{\sigma^{2^n}_y}{(m-1)})$ with $\sigma^{2^i}_y \sim \chi^2(m)$ and m being the number of samples drawn. The correlation coefficients were drawn randomly from a uniform distribution $CC_y \in \mathbb{R}^{n \times n}, cc^{ij}_y \sim \mathcal{U}(0, 1)$. With $\circ$ denoting the element-wise multiplication (Hadamard-product) the covariance matrix was then calculated as

$$\boldsymbol{\Sigma}_y = (CC_y \circ A) \circ (\boldsymbol{\sigma}^T_y \boldsymbol{\sigma}'_y). \tag{42}$$

This procedure does not have to result in a positive semi-definite matrix, a characteristic required for covariance matrices. To ensure that the eigenvalues are positive I rescaled the non-diagonal elements such that $\sum_{i \neq j}(|\sigma_{ij}|) = \sigma_{ii}$, with entries σ_{ij} for Σ (Gershgorin circle theorem [74]).

Starting from the noiseless sample set $\mathbf{T}_y$ noisy sample sets $\mathbf{T}_y^{\sigma_n}$ were generated in analogy to Zhang et al. [251]. Hereby, each sample $\mathbf{x} \in \mathbf{T}_y$ was modified as $\mathbf{x}^{\sigma_n} = \mathbf{x} \cdot \boldsymbol{\sigma}_{\mathbf{mult}} + \boldsymbol{\sigma}_{\mathbf{add}}$, with multiplicative noise $\boldsymbol{\sigma}_{\mathbf{mult}} \sim \mathcal{N}(0, \sigma_n)$ and additive noise $\boldsymbol{\sigma}_{\mathbf{add}} \sim \mathcal{N}(0, 1)$.

3.3.1.3 Parameter settings

The initial parameters were set based on Zhang et al. [251]. For a prototype in Equation 39 without predecessors the parameters were $\mu_\mu = 6$ and $\sigma_\mu = 1.5$ and $\mu_\sigma = 28.11$ and $\sigma_\sigma = 44.25$. The α and β parameter of the Γ distribution (Equation 40) were set as 10, as this leads to a fold change of around 1 for most of the genes but also allows for fold changes higher than 2, which is a reasonable assumption for gene expression datasets.
To estimate the number of interactions within the Barabasi-Albert graph I used the ratio that is calculated based on the amount of proteins and the amount of interactions in the human protein network. As the estimated size of the human interactome is ≈650.000 [198] and the proposed number of proteins ≈19.000 [63], I considered to have $30n$ interactions.
The noise level in $\mathbf{T}_y^{\sigma_n}$ varied from $\sigma_n = (0, 0.05, \ldots, 0.5)$. I tested dimensions of $n = (2, 10, 50, 100, 500, 1000)$, chose $\sigma_p = 1.5$ based on Zhang et al. [251] and used $n = 1000$ for the final simulations. I set $m = 10$, because of the low sample number in the real cancer data, the model should fit to.

For each global structure I sampled 100 datasets. The analysed structures can be divided into three groups: Patterns for which I did not impose any relation, patterns with linear relations and patterns with tree-like structures. The patterns are given in Figure 19 and the class dependencies were imposed as pairwise relations, as for example for the linear pattern with 5 nodes: $\mathbf{O} = \{y_0 \prec y_1, y_1 \prec y_2, y_2 \prec y_3, y_3 \prec y_4\}$
For the structure of tree_N6(4) I used an additional prototype $\mathbf{p_a}$ and symmetric changes to construct the merging structure. It holds that $\mathbf{p_0} = \mathbf{p_a} + \mathbf{r_0} \circ \mathbf{f_0}$ and $\mathbf{p_1} = \mathbf{p_a} + \mathbf{r_1} \circ \mathbf{f_1}$. The next prototype is $\mathbf{p_2} = \mathbf{p_0} + \mathbf{r_1} \circ \mathbf{f_1}$. Afterwards, the prototype $\mathbf{p_a}$ was deleted from the dataset. In this way I sampled an inverse diverging structure without explicitly inverting the relations between $y_0 \prec y_2$ and $y_0 \prec y_1$.

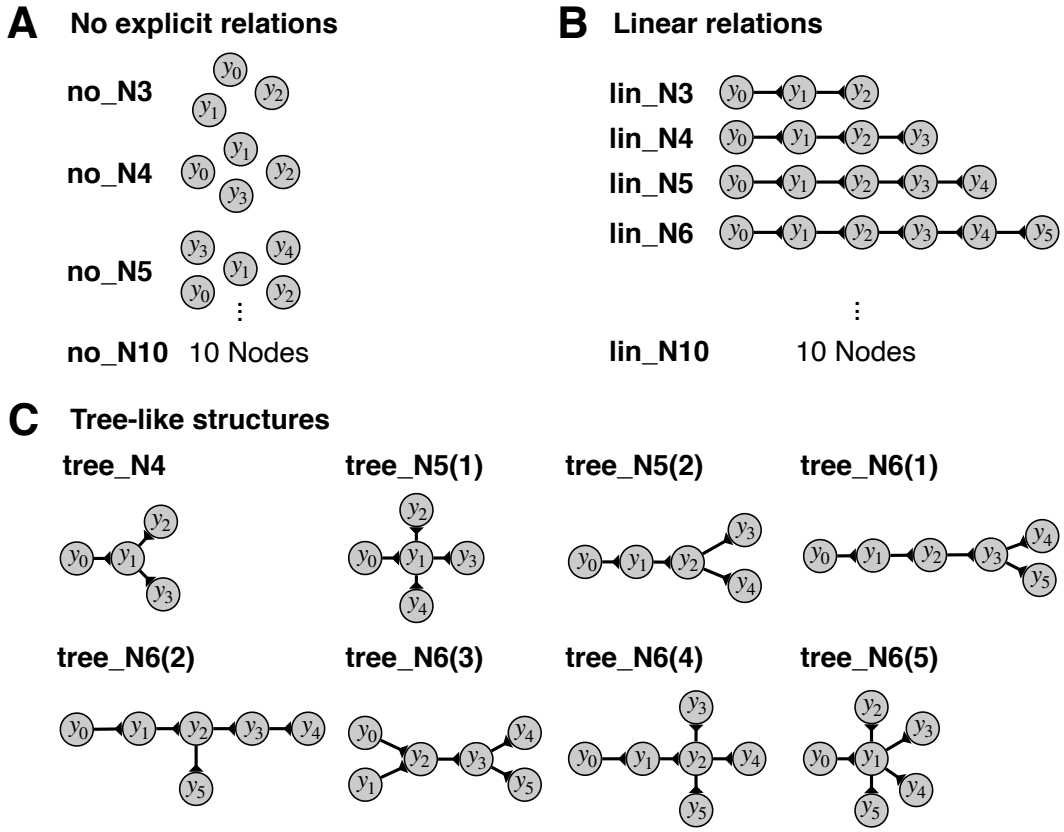


Figure 19: Patterns of ordinal relations underlying the sampled artificial data. The sampled patterns can be divided into three groups. A: The first group (no) comprises datasets on that no explicit relation is imposed. These datasets are made up of 3-10 classes. B: The second group (lin) comprises linear relations. C: The third group (tree) are tree-like structures. Here all trees up to 6 nodes are sampled. The notation of each dataset is made up of the group (no, lin, tree), the number of nodes (N) and in case of the tree-like structures an additional counter.

3.3.2 Characterisation of the artificial data

3.3.2.1 Likelihood of sampling circles

As the distances between the prototypes of my artificial data were drawn randomly, it might happen that circles were sampled by chance, which do not fit the expected behaviour. A circle might be considered as a new prototype being nearby a previous prototype. As the step size between two prototypes was not normalised in this model, it depends on the dimensionality. The higher the dimensionality the larger the distance between neighboured prototypes. The prototype sampling using the linear relation patterns and 100 repetitions resulted in median

pairwise distances of direct related prototypes of $(1.4, 3.3, 7.4, 10.5, 23.4, 33.2)$ for $n = (2, 10, 50, 100, 500, 1000)$ (Appendix Figure A1).
The sampling procedure could be compared to a random walk and is a random walk if the step sizes in sampling would be normalised to 1 ($|\mathbf{r} \circ \mathbf{f}| = 1$ in Equation 40). In this case the distances of neighboured prototypes do not depend on the dimensionality anymore.

Analysing prototype distances under a random walk assumption (normalised step size) and in the model (unnormalised step size) it can be seen that both show the same behaviour: the pairwise distances increased with the steps performed and they were less distributed the higher the dimensionality (Figure 20).

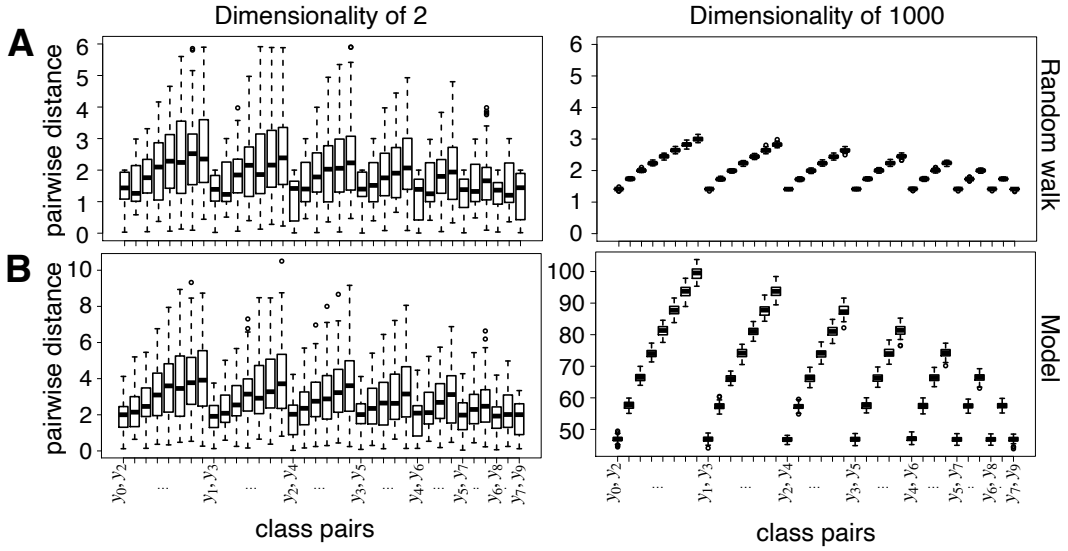


Figure 20: Pairwise prototype distances of linear datasets with 10 classes. Distances of not directly neighboured classes using A: a normalised step size and B: an unnormalised step size for a dimensionality of 2 (left) and 1000 (right). There were 100 sampling repetitions for this pattern performed.

Using this sampling experiment I measured the frequency of prototypes that are not neighboured by construction being close to another prototype. I based the measure of being nearby on the step size of the model which relates to the distance of two prototypes that are neighboured by construction. Under the random walk assumption this corresponds to a distance of 1. Within the biological related model I used the 75 % and the 95 % quantile of the distance distribution of neighboured prototypes. It could be seen that the frequency of potential circles decreased with the dimensionality in all cases and decreased to zero for the patterns with an assumed linear relation and without any relation. In case of the random walk assumption and a distance of 1 the frequency returned to zero for all analysed

patterns. This was also the case for the model, with the exception of one dataset, the dataset that has two roots (Figure 21). This dataset was included to model all trees of size 6 but does not fit to the relational one predecessor assumption. The two root prototypes are likely to be seen as neighboured, although they are not directly related based on the construction method.

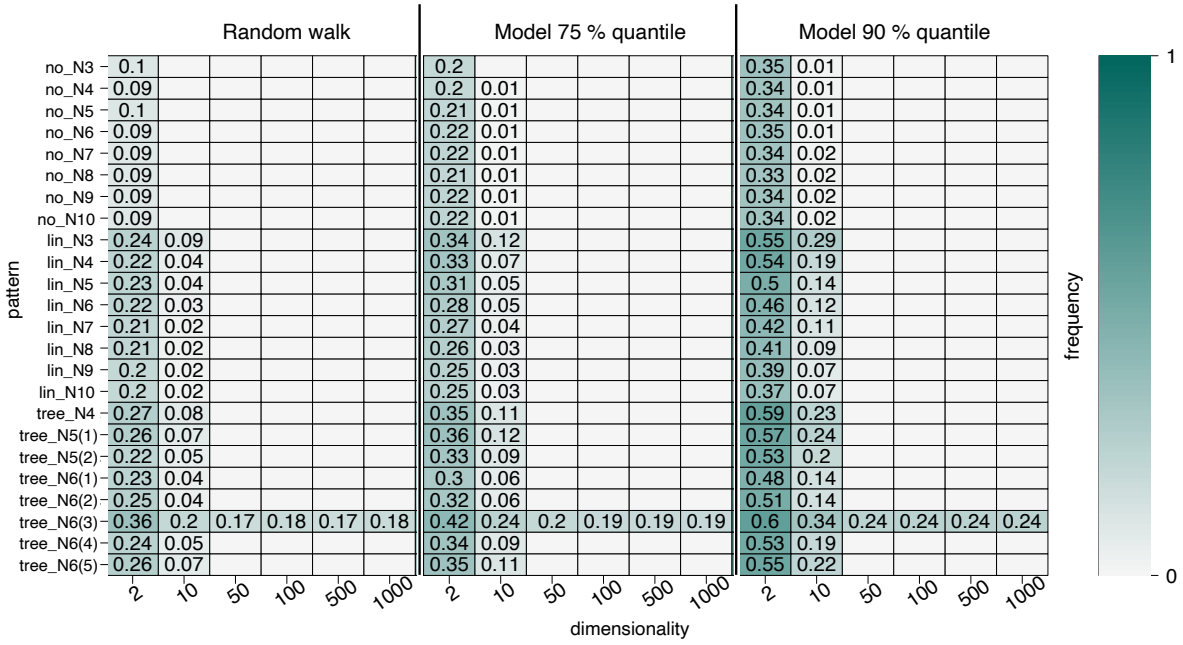


Figure 21: Probability to sample a circle on the prototype level. As for the random walk the step size is 1, the probability is given as the amount of prototypes that show a pairwise distance smaller than 1. For the model the border characterising the sampling of a circle is given once by the 75 % quantile and once by the 90 % quantile of all step sizes in the sampling procedure. As tree_N6(3) has two starting nodes those are located nearby.

The analysis of the frequency of samples being nearby a prototype that is not their prototype showed that this behaviour is similar using the biological related model but not using the random walk assumption and a normalised step size. As reference distance on the sample level the 75 % and the 95 % quantile of the distances of samples towards their own prototypes was used (Appendix Figure A2).

As consequence the likelihood to sample a circle can be assumed as zero in the proposed biological related model and must not be further considered.

3.3.2.2 Distance comparison with real data

The effect of an increase in pairwise distance between prototypes with the number of steps (Figure 20) was only transmitted to the samples using not the random walk assumption and hence an unnormalised step size. Furthermore, this effect resembled the behaviour seen in expression data sampled at different points in time during *Caenorhabditis elegans* (*C. elegans*) development (Figure 22).

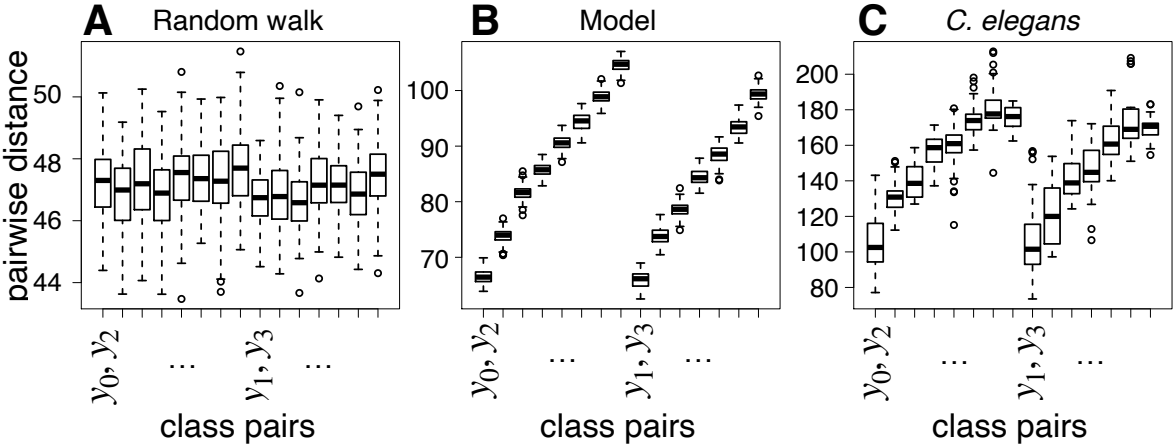


Figure 22: Pairwise sample distances of linear datasets with 10 classes. The figure shows the distances of not directly neighboured classes of class y_0 and y_1 of a sampling using A: normalised prototype step size and B: unnormalised prototype steps and a dimensionality of 1000. The pattern was sampled 100 times. C: Additionally, the pairwise distances between expression data of presumable not directly neighboured classes of class y_0 and y_1 sampled at 10 different points in time during *Caenorhabditis elegans* (*C. elegans*) development (d_6) are shown.

3.3.3 Characteristics of the set of returned cascades

I analysed this artificial data for different levels of multiplicative noise using the Occ with a linear support vector machine (SVM) (L2-SVM) as base classifier.
One prerequisite of finding any cascade is a good pairwise separability. The analysis of all pairwise class-wise separabilities revealed a perfect separation for small noise levels. Unexpectedly, if the noise level increased the separability decreased (Figure 23).

3.3.3.1 Subcascades still carry information

Analysing the returned cascades showed a similar behaviour. For small noise levels only the expected cascades were returned. Whereas if the noise level increased, the distribution of subcascades spread. In the case of a linear relation of seven classes

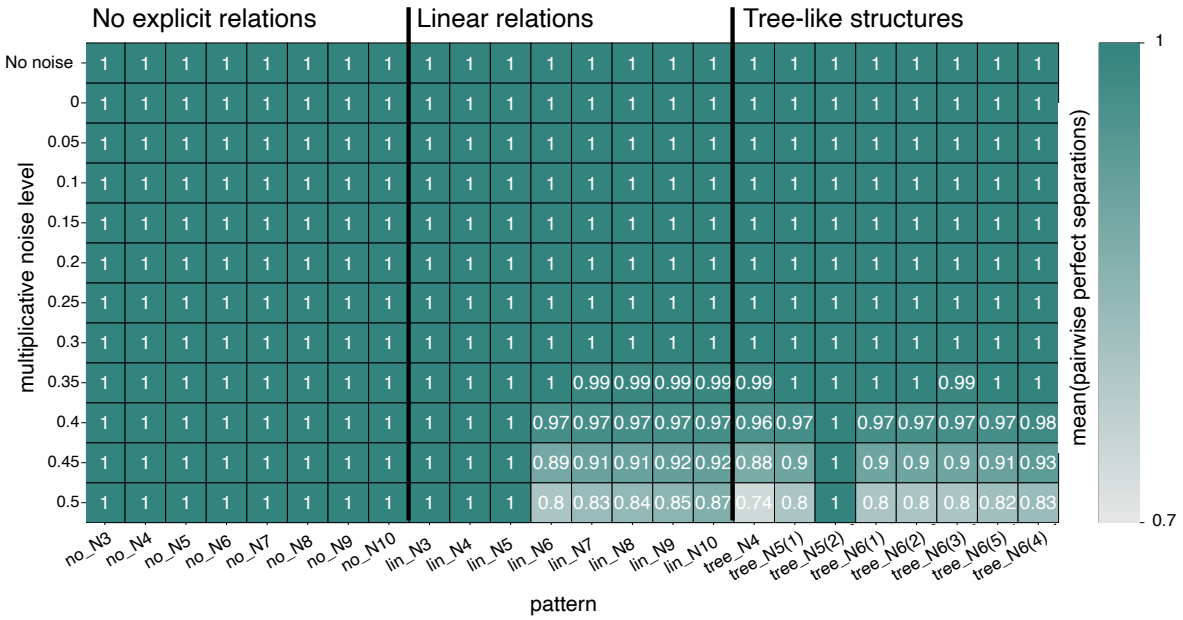


Figure 23: Pairwise class separability with increasing levels of noise for all sampled patterns. The average percentage of perfect pairwise separations for all patterns and all levels of multiplicative noise are shown. The underlying data had a dimensionality of $n = 1000$ and there were 100 datasets sampled per pattern.

for example, there were no longer two cascades of length 7 returned but several cascades of length six (Figure 24).
Nevertheless, even if the total cascades were not returned due to noise, the returned subcascades could give insights about the relations. One possibility to filter the returned cascades is by setting a threshold on the minimal class-wise sensitivity. I always required a minimal class-wise sensitivity of 70 % to gain a good classification result. As smaller cascades can achieve higher sensitivities easier (Theorem 2), I analysed different sets of returned cascades that pass the minimal class-wise sensitivity: all cascades longer or equal three, the longest cascades and the Pareto optimal set.

3.3.3.2 Swap of the last two classes as possible error type

For the artificial data I know which cascades I expect and hence I checked the average frequency of not expected cascades in the returned sets, as $\mathrm{NEDR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TP}}$, with FP not expected but found and TP expected and found (Figure 25). The corresponding standard deviations can be found in Appendix Figure A3.
Comparing the set of longest cascades passing the threshold with the set of all cascades passing the threshold, the first set showed on average lower error rates

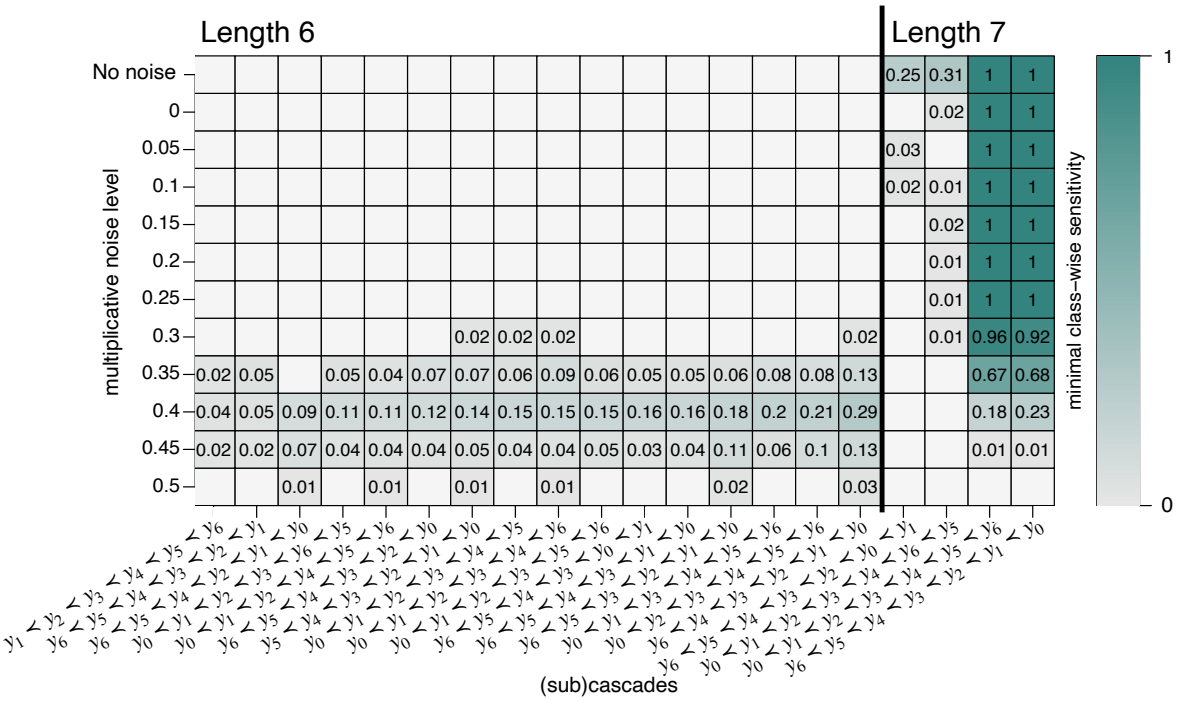


Figure 24: Returned cascades for the linear pattern with 7 classes. The cascades of length 6 and 7 and their minimal class-wise sensitivities, using samples of dimensionality 1000, are shown. It can be observed that the cascades of length 7 break apart with an increasing noise level (y-axis). Nevertheless, the different parts are subcascades of the cascades of length 7 and hence still carry information about the original order. There were 100 datasets sampled.

for linear related patterns. The error in the returned cascades observed for this sets is a swap in the last two classes. Including the subcascades seemed to reduce the amount of not expected cascades for the not linear relations but to increase the amount for the linear relations.
Finally, the Pareto set was analysed. This corresponded here in many cases to the best longest cascades as the minimal class-wise sensitivities of the longest and the subcascades equaled 1. However, especially for high noise levels, also subcascades were included. The Pareto optimal set reduced on average the amount of not expected cascades compared to the set of longest cascades with a global threshold, particularly for low noise levels. Compared to the set of all subcascades the Pareto set performed on average better for linear patterns and similar for tree-like patterns.

Analysing the set without any imposed relation pattern each returned cascade is not expected. The average not expected positive rate across the noise levels $\mathrm{NEPR} = \frac{\mathrm{FP}}{\mathrm{N}}$, with N being the overall number of possible permutations of a given subcascade size, showed a dependency on the number of classes and the subcas-

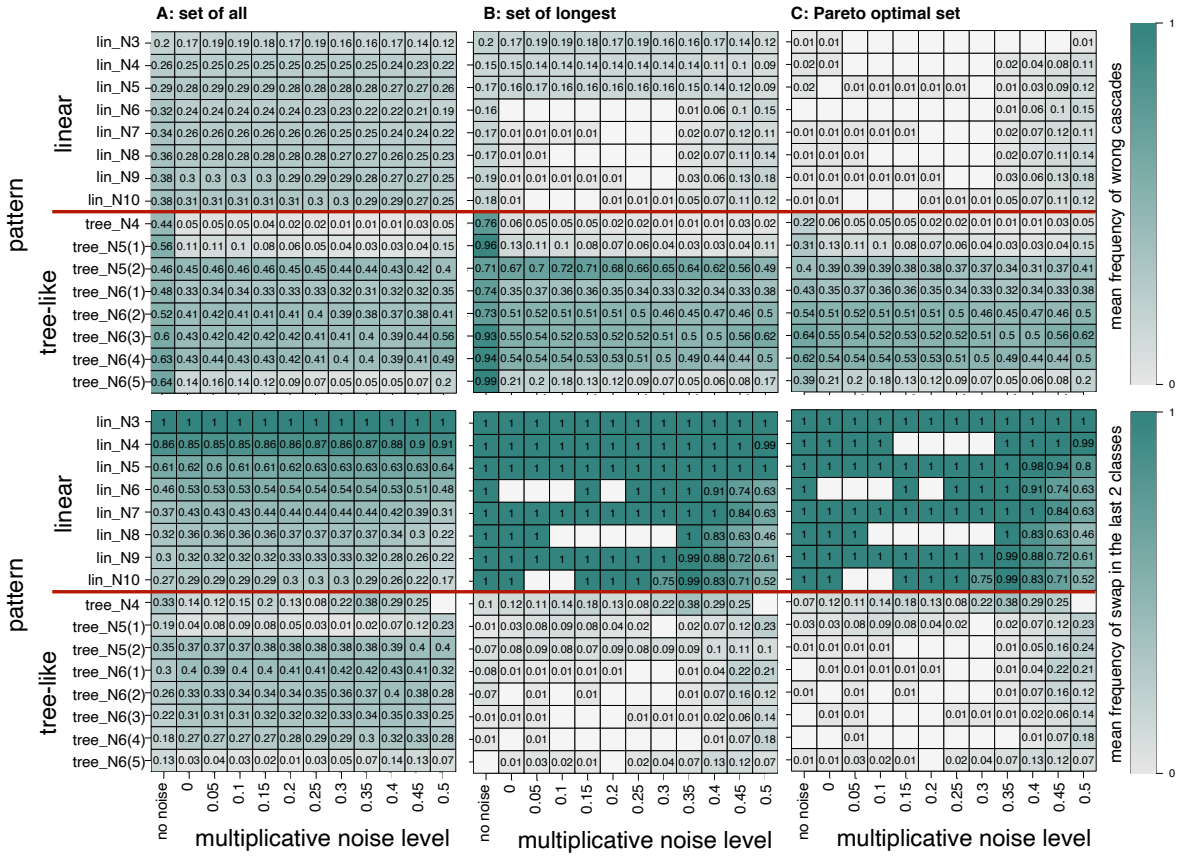


Figure 25: Mean frequency of not expected cascades for all patterns. A global threshold of 0.7 was imposed on the screening results. Three different sets of remaining cascades were compared to the set of expected cascades that is given by the underlying corresponding pattern (upper part). A: First all top ranked cascades of all lengths are used. B: Second, only the longest once passing the threshold are shown. C: Third, the Pareto optimal set is given. Additionally, the frequency of an often occurring error is shown (lower part), which is a swap in the last two classes of a cascade. The underlying data had a dimensionality of $n = 1000$ and there were 100 datasets sampled per pattern.

cade size. Whereas it dropped beyond 1 % for datasets comprising more than 4 classes and cascades longer than 3, independent of the analysed set (all, longest, Pareto), it ranged from $3 - 36$ % for cascades of length 3 for all number of classes. For cascades of length 4 in a dataset comprising 4 classes it was around $1 - 2$ %. Already for cascades of length 5 the NEPR was in the range of 10^{-5} (Table 4).
Generally it could be observed that cascades of length 3 are more likely to be false positive findings and that analysing the Pareto optimal set taking subcascades into account might have a beneficial effect on the frequency of not expected cascades.

Table 4: Rate of found not expected cascades (NEPR) within the set of patterns without relations. For all three selection methods the datasets without any imposed relational patterns were analysed and the average false positive rate across all noise levels is given in percent. For subcascades of length 6 or longer the false positive rate is summed up and shown. All analysed cascades were filtered for passing a minimal sensitivity threshold of 0.7. The set of all comprises furthermore all cascades of length longer or equal 3, the set of longest comprises only the longest cascades returned and the Pareto optimal set is calculated for length $\geqslant 3$ and minimal class-wise sensitivity $\geqslant 0.7$. The underlying data had a dimensionality of $n = 1000$ and there were 100 datasets sampled per pattern.

	set of all				set of longest				Pareto optimal set			
length / pattern	3	4	5	$\geqslant 6$	3	4	5	$\geqslant 6$	3	4	5	$\geqslant 6$
no_N3	35.90	-	-	-	35.90	-	-	-	22.82	-	-	-
no_N4	36.23	1.60	-	-	25.16	1.60	-	-	11.92	1.27	-	-
no_N5	11.69	0.41	0.01	-	7.31	0.35	0.01	-	8.43	0.17	0.01	-
no_N6	12.16	0.40	0.01	0.00	7.06	0.23	0.01	0.00	7.79	0.10	0.00	0.00
no_N7	11.46	0.39	0.01	0.00	5.76	0.15	0.00	0.00	6.01	0.08	0.00	0.00
no_N8	11.84	0.39	0.00	0.00	4.72	0.11	0.00	0.00	4.79	0.07	0.00	0.00
no_N9	11.87	0.40	0.01	0.00	3.78	0.05	0.00	0.00	3.81	0.08	0.00	0.00
no_N10	11.78	0.41	0.01	0.00	3.03	0.04	0.00	0.00	3.03	0.07	0.00	0.00

For all datasets the set of returned cascades contained the expected relations. These cascades might be analysed as sequences separately, but they might also be aligned to a graph structure. Next, I analysed the power of the cascades algorithm to reconstruct the original graph pattern and compared this technique to other graph reconstruction techniques.

3.4 Phenotype network reconstruction

I did not only analyse the power on reconstructing the relational patterns of the cascades algorithm using a linear SVM as base classifier but also its suitability in reconstructing phenotype networks. I, furthermore, compared it to additional methods that might be used to tackle this task: a supervised and an unsupervised version of self-organising maps (SOMs) (Section 2.3.2), different distance measures, like correlation and mutual information (MI) (Section 2.3.1) and the Occ with various base classifiers (SVM with different kernels, random forest (rF), k-nearest neighbour (kNN)). I performed leave-one-out-cross-validation (LOOCV) experiments and a parameter tuning optimising for accuracy for the number of trees in the case of rF, k in the case of kNN and the cost parameter in the case of the SVM approaches.

Furthermore, I included a linear L2-SVM with a fixed cost parameter for all base classifiers as 1 and a heterogeneous base classifier ensemble. Within this heterogeneous ensemble for each binary classifier the classifier with the best performance in terms of accuracy among all types of base classifiers was selected.

3.4.1 Reconstruction procedure

Depending on the type of algorithm output a weighted adjacency matrix was calculated (Figure 26).

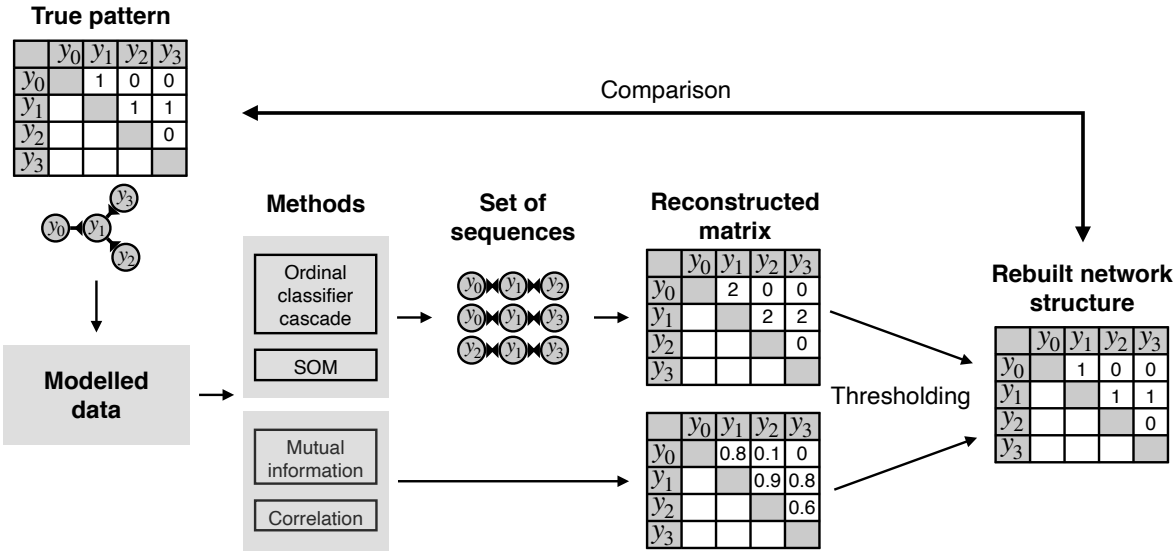


Figure 26: Scheme on the network reconstruction procedure. The relational pattern can be expressed as a graph structure and its corresponding adjacency matrix. This matrix is used as ground truth and used to generate data following the given pattern. The data is analysed using a set of methods, the ordinal classifier cascade (Occ) and self-organising maps (SOMs) as well as the pairwise mutual information and correlation. The former result in a set of cascades, which are aligned to a graph structure. The later result directly in an adjacency matrix. A thresholding procedure results in unweighted graphs, which are consequently compared to the ground truth network.

The cascades algorithm as well as the SOM screenings result in sets of returned sequences. To reconstruct a graph structure, described by its adjacency matrix A from these sequences, I filtered the set of returned cascades imposing high performance criteria. I used the Pareto set and a threshold of 1. I counted the occurrence of each relation and normalised with the number of cascades in the set. This resulted in the adjacency matrix A.

Calculating the average pairwise correlation and normalised mutual information directly an adjacency matrix A was generated.

By imposing a threshold on the edge weights of A an unweighted adjacency matrix was generated and the effect of this threshold was tested (Section 3.4.3). The reconstruction performance was evaluated based on the edge error rate $\text{err} = \frac{\text{FP}+\text{FN}}{\text{PE}+\text{NE}}$ of the unweighted graph with respect to the known true graph pattern, with FP denoting the number of found edges if an edge existed, FN as the number of not found edges, although there is an edge, PE as the number of edges and NE the number of possible edges that do not exist.

3.4.2 Impact of the bin size on the mutual information

The final error rate of the the mutual information methods depended on the bin size used to discretise the data (Figure 27). I discretised the data into 5 bins, as this showed the lowest error rate.

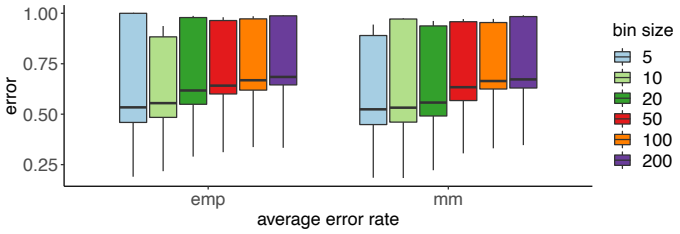


Figure 27: Dependency of the error rate on the bin size using mutual information as dependency measure. Error rate distributions of the two mutual information (MI) measures and all patterns using different bin sizes and data without noise. The underlying data had a dimensionality of $n = 1000$ and there were 100 datasets sampled per pattern. emp=empirical and mm=Miller-Madow corrected MI.

3.4.3 Global and dataset specific thresholding

I imposed two different kinds of thresholding in order to result in an unweighted graph: a global and a dataset specific thresholding. To evaluate the impact of the global threshold I screened through a range of thresholds $(0, 0.1, \ldots, 1)$ using the data without any noise and reported the average error rate per group of pattern (Figure 28). It could be seen that there is a strong impact of a global threshold on the performance of the dependency methods and the group of no relations, as mutual information and correlation calculation rarely resulted in a value of 0, directly. Also, for the SOMs a small threshold reduced the average error rate

to 0. The beneficial impact of a global threshold on the performance of the SOM on datasets with no relational pattern was counteracted by its impact on datasets with relations, as within these sets the error rate increased as soon as there was any global threshold. This was not the case for the dependency methods, which indicated a slight improvement also in these groups of patterns.
The Occs showed a high variance depending on the type of base classifier. The variance was low for the SVM base classifiers independent of the kernel, except for the linear L1-SVM. In general it could be seen that the Occs always performed better in reconstructing relations independent of the global threshold.

The dataset specific threshold was set by screening over the range of the values in the adjacency matrices with a step size of 0.05 and checking for the maximal threshold at which the resulting graph is still connected. It is used as refinement on potential indirect relations.

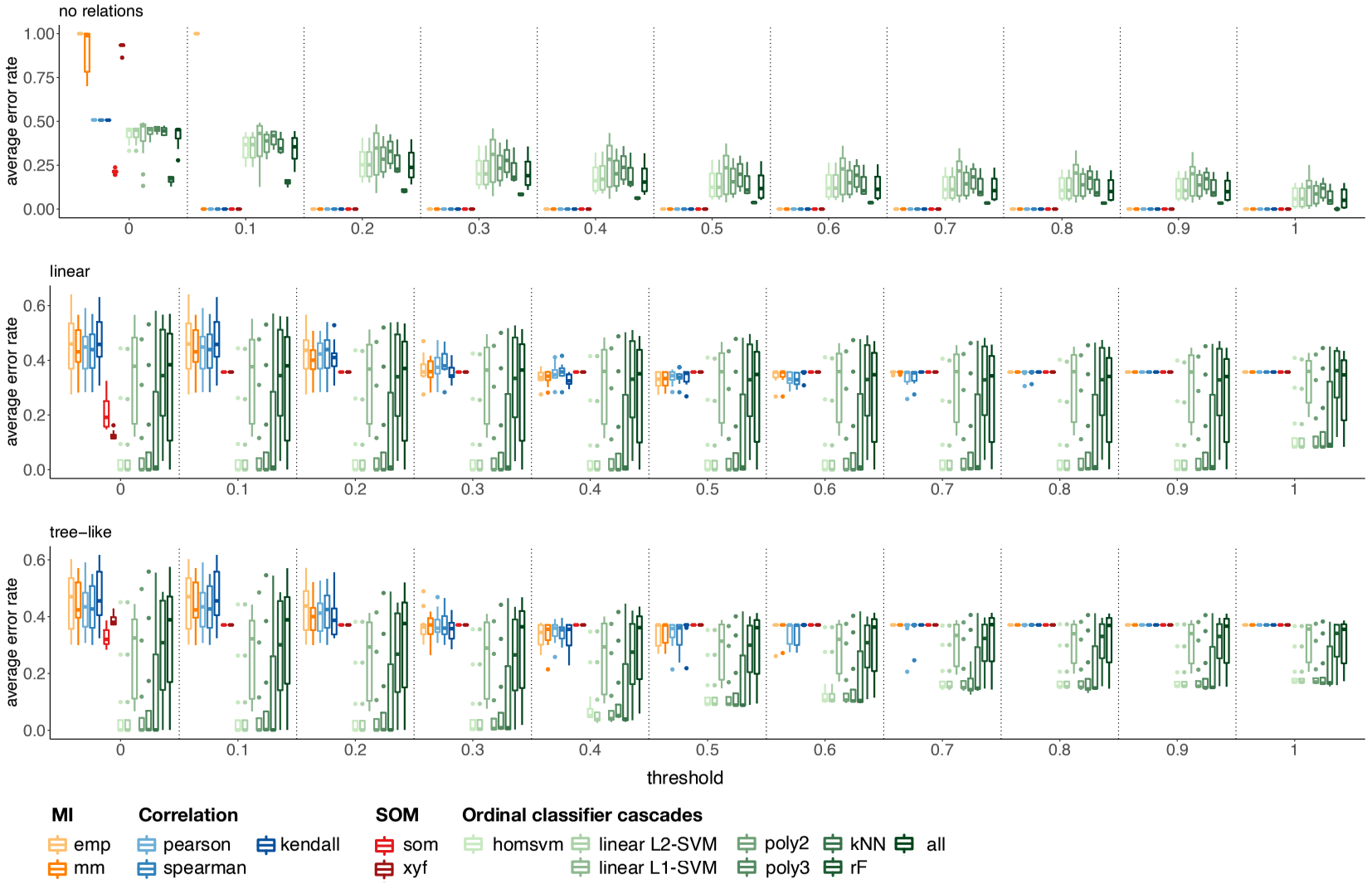


Figure 28: Average error rate per pattern group and method dependent on the global threshold. (Continued on the following page.)

Figure 28: (Previous page.) Average error rate per pattern group and method dependent on the global threshold. During the network reconstruction various global thresholds on the frequency of edge occurrence are tested using the dataset without any noise. The thresholds (x-axis) ranged from $0-1$. The individual methods are grouped and colour coded: MI=Mutual information (emp=empirical, mm=Miller-Madow corrected), Correlation (spearman, pearson, kendall), SOM=self-organising map (som= unsupervised version, xyf=supervised version) and ordinal classifier cascade (Occ) (homsvm=linear L2-support vector machine (SVM) with constant cost parameter for all base classifiers; linear L2-SVM and linear L1-SVM, SVM with polynomial degree of 2 and 3 (poly2 and poly3), k-nearest neighbour (kNN), random forest (rF) with parameters optimised in a hyperparameter tuning; heterogenous, which is an ensemble selecting the best binary classifier per step from the set of parameter tuned classifiers). The underlying data has a dimensionality of $n = 1000$ and there were 100 datasets sampled per pattern.

3.4.4 Reconstruction results

The final reconstruction results showed the performance differences between the pattern group of no relations and any relations. The results without any noise and with a noise level of 0.25 and a global threshold of 0.3 are shown in Figure 29. The results for different strengths of multiplicative noise are given in the Appendix Figure A4. The mutual information performed worst for all patterns. Choosing a very high global threshold this performance could be largely improved for the group of no relations and slightly for the other groups, but it would still not be better than the other methods (Figure 28). The SOMs showed an increasing performance with the number of classes and no affection by the noise level. Instead they resulted in exactly those error rates that would have been assumed for random guessing. Already by construction, they are not suitable for tree-like structures, as they try to adapt a symmetric grid structure to the data, but the results showed that they are in general not suitable for finding ordinal relation patterns.
The Occs revealed a performance difference for the group of no and with relations and between feature selecting base classifiers and base classifiers without feature selection. Only the two base classifiers with feature selection and the heterogenous cascade became worse with an increasing noise level in the groups of patterns with relations. The rF performed best if no relation was imposed. The remaining base classifiers still showed a performance increase with the number of classes. The classifier cascades built upon a SVM (exception linear L1-SVM) and kNN could almost perfectly reconstruct the ordinal relational patterns using these settings.

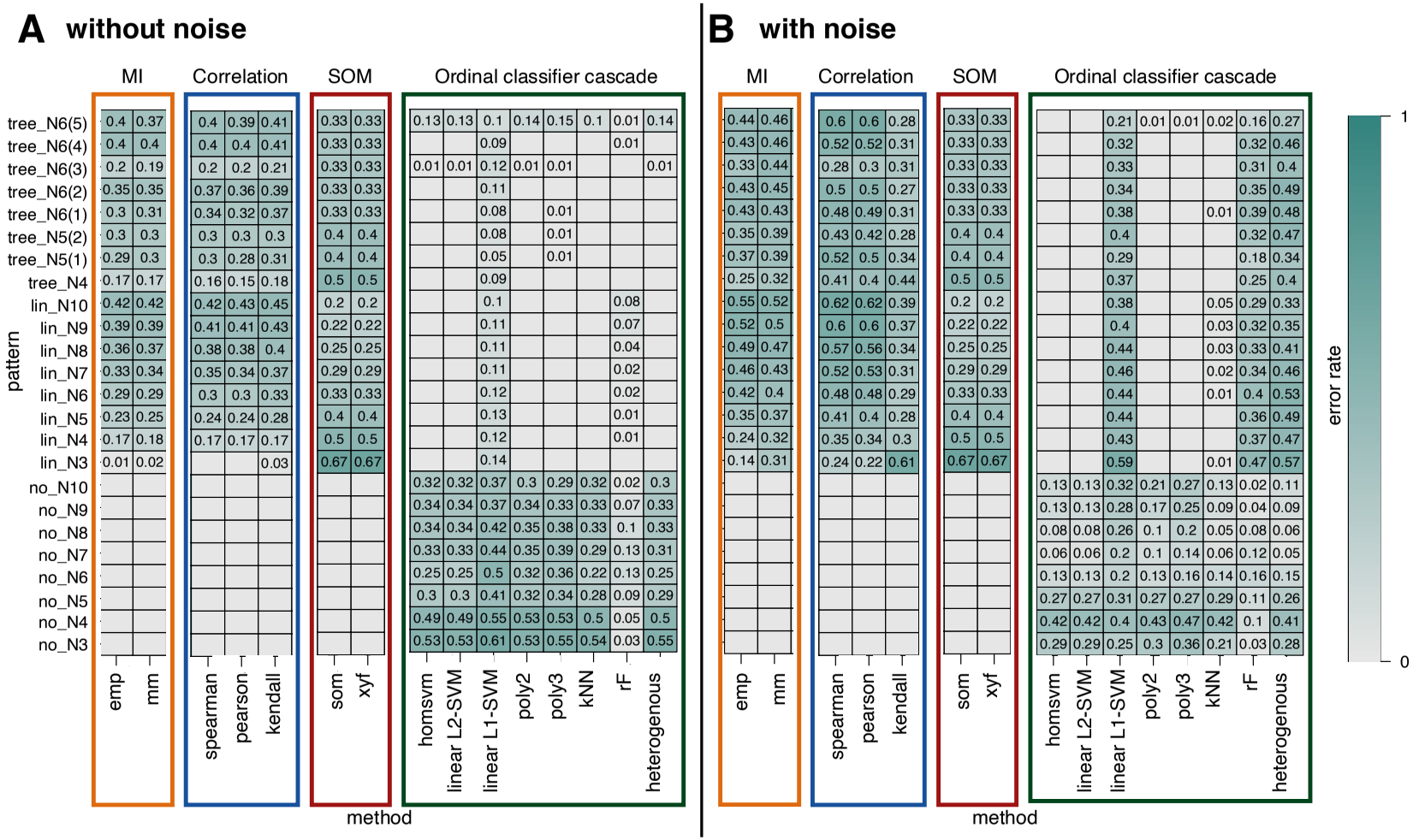


Figure 29: Error rates of reconstructing patterns using a global threshold of 0.3. A: The left panel shows the average error rates per pattern and method if the set of longest cascades with a minimal class-wise sensitivity of 1 is analysed and no noise is added. B: The right panel shows the results of a noise level of 0.25. (Continued on the following page.)

Figure 29: (Previous page.) Error rates of reconstructing patterns using a global threshold of 0.3. A: The left panel shows the average error rates per pattern and method if the set of longest cascades with a minimal class-wise sensitivity of 1 is analysed and no noise is added. B: The right panel shows the results of a noise level of 0.25. All cascades passing a threshold of 1 are used for reconstruction. The methods are grouped: MI=Mutual information (emp=empirical and mm=Miller-Madow corrected), Correlation (spearman, pearson, kendall), SOM=self-organising map (som= unsupervised version, xyf=supervised version) and ordinal classifier cascade (Occ) (homsvm=linear L2-support vector machine (SVM) with constant cost parameter for all base classifiers and linear L2-SVM, linear L1-SVM, SVM with polynomial degree of 2 and 3, as poly2 and poly3, k-nearest neighbour (kNN), random forest (rF) with parameters optimised in a hyperparameter tuning and heterogenous, which is an ensemble selecting the best binary classifier per step from the set of parameter tuned classifiers). The underlying data has a dimensionality of $n = 1000$ and there were 100 datasets sampled per pattern.

3.5 Discussion and conclusion

A fast and efficient calculation procedure is required to screen for ordinality in high-dimensional datasets with many classes. The proposed preselection strategy in combination with the prediction map look-up table structure provides an adequate solution for this task. The benefit of the preselection strategy depends on the search and dataset characteristics. Theorem 1 and Theorem 2 clarify that the number of (sub)cascades that are discarded depends on a combination of the length of the cascade that is searched for and the given threshold. The possibility to filter out orderings from the search space due to Theorem 1 and the reduced number of remaining possible orderings for evaluation was shown only for the Occ. The prediction map, however, might also be easily employed by other classifier schemes that are based on binary decomposition. Combining the screening core functions with functions helpful for analysing results in an R-package *ORION* makes the methodology publicly available in an accessible way.

The general applicability of the Occ to reveal relational sequences was tested and characterised by the use of artificial data, modelled to resemble relational processes between gene expression stages (Section 3.3.1). I used data of total ordinal relations, partial ordinal relations and without any relational pattern imposed.
I furthermore proposed a strategy to use the set of relational sequences to reconstruct the original pattern and compared it to other methods, mutual information and correlation, which have been employed for reconstructing gene-gene or protein-protein interactions [197, 254], but not on the phenotype level. In order to get insight about the impact of the choice of base classifier in reconstruction I performed experiments using a common set of binary classifier constructs. Additionally, the SOM was tested on their capacity to reveal ordinal patterns albeit its applicability on partial orders is not expected due to the linear grid imposed.
The experiments showed that the Occ using a linear L2-SVM is able to find relational patterns between all classes but also between subset of classes. One aspect that influences the performance is the pairwise separability. If there is a relation between all classes but the pairwise separability is not given, cascades between subsets are returned and still show a general trend (Figure 24). The inclusion of too many subcascades passing a specific global threshold within the set of returned cascades for further analysis is however not always advisable. This was experimentally shown (Figure 25) and theoretically supported by the fact that shorter cascades perform at least as good as their prolongation (Theorem 2) and hence reach a specific threshold easier. To overcome this the Pareto optimal set, comprising the best solutions of all lengths, might be used. This selection showed a beneficial effect on the total linear orders. The beneficial effect in case

of the linear patterns is hereby rather due to the exclusion of longest cascades with a lower performance than due to the inclusion of smaller cascades, as often the best performing longest cascades show a minimal class-wise sensitivity of 1. For the partial orders it performed at least as good as the set of longest cascades. Nevertheless, the set of all cascades most often showed the least amount of not expected cascades, especially if the frequent error of a swap of the last two classes is additionally considered and excluded. The performance might be improved by higher thresholds, similar to the effect of the change from a threshold of 0.7 for the set of longest cascades to the Pareto optimal set, which corresponds here in most cases to the longest cascades with a threshold of 1.
Analysing this swap it can be seen that if the set of longest cascades contains an erroneous cascade the error is, except for high noise levels, always a swap in the last two classes (Figure 25). Conclusively, this swap is most likely not dataset specific but can be referred to the less amount of constraints at the end of the classifier cascade. The analysis on the artificial data revealed additionally that cascades of length 3 are more likely to represent false positives (Table 4) and should be handled with care.

The analysis of the capacity to reconstruct relational patterns showed a good performance of the Occ with respect to the other methods under investigation and independent of the base classifier used (Figure 29). However, the feature selecting base classifiers seem to perform worse in the patterns that include relations and are more sensitive to noise. As by construction all features tend to be informative within this set of patterns, this observation is most likely due to the dimensionality reduction and a general conclusion on a disadvantageous effect of feature selection on the reconstruction capacity cannot be drawn.
A high threshold of 1 and all cascades longer or equal length 3 are included resulting in many not expected subcascades in case of no relational pattern. The more classes a dataset comprises the less not expected relations are revealed, as these mainly result from cascades of length 3 or 4 (Figure 4). This indicates that within the analysed dimension the longer the returned cascades, the less likely they represent false positives. This observation is supported by the decomposition of the Occ prediction rate, as for long cascades there are more external rates evaluated counteracting the curse of dimensionality. Strikingly, the rF as base classifier produces much less false positives within this set of no explicit relations than all other base classifiers, with the linear L1-SVM producing on average more false positives. The feature selecting strategy seems on the one hand beneficial and on the other hand disadvantageous within the set of no relations.
Nevertheless, within the set of analysed methods the Occ works, due to the false findings of cascades of length 3 and 4, worst in showing and finding absent

relations as soon as there is any minimal threshold on the edge weights. But its performance on this set can be improved by higher edge weight thresholds.
Without noise the correlation and mutual information measures perform similar. The error rate increases with the number of classes, most likely due to measuring more transitive edges. Using a high noise level all methods end up in showing almost the same error rates as random guessing.
The SOM shows the inverse behaviour. The more classes the better the performance. It is neither adopted to any noise level, nor does the performance depend on the type of learning (supervised/unsupervised). It shows exactly those error rates that are expected for random guessing and is completely unsuitable for the given task.

Using artificial data I could show that the Occ can be used to find relational sequences within a set of classes and subsets of classes. Furthermore, the sequences found in subsets of classes can be used to reconstruct a neighbourhood graph using the proposed procedure. This reconstruction works better than using MI or correlation, most likely because Occ does not reveal pairwise relations but overarching patterns taking all degrees of neighbourhood into account for one (sub)cascade.

One might argue that the Occs relies on the separability of classes and hence the chances to find cascades increases with the dimension.
Separating categories within high-dimensional data the curse of dimensionality has to be considered in order to not produce artefacts. In contrast to a common multi-class classifier system (OaO), the Occ seems to be unaffected by this. The experiments showed that the number of cascades found does not increase with dimensionality for dimensions up to 10000 (Figure 18). For the same setting an OaO classifier system would always find a perfect solution. The observed behaviour could be ascribed to the usage of external classes, classes that are not used in training, but for evaluation. To investigate potential relations within the biological and medical context, this characteristic might be highly beneficial as by the use of external classes for example no feature selection and hence no preselection of potential target genes has to be performed to work against the curse of dimensionality.

4

CHARACTERISATION OF ORDER TYPES

In the previous section it could be observed that the type of the base classifier influences the reconstruction performance. As different types of base classifiers are suitable for different tasks, I asked whether the characteristics of binary classifiers can be used to characterise the type of ordinality detected.

An ordinal relation in 1D is intuitive as there is a one-to-one relation between the ordinality in feature space and the ordinal relation within the label space. For high dimensions different types of class patterns might be considered as ordinal (Figure 8).
I first defined different characteristics that can be used to distinguish different types of ordinality and then used the ordinal classifier cascade (Occ) with different base classifiers on artificial data to show that one can distinguish between some criteria based on the classifier results.
The focus of my study is the difference between local and global ordinal structures distinguished by base classifiers that allow for disconnected and solely connected decision regions and to point out that base classifiers with disconnected decision regions bear the risk of reporting representations as ordinal, although they are not considered as ordinal, like a partial repetition of an ordinal pattern [120].

4.1 Differentiation of types of ordinality

I defined four different criteria of ordinal concepts in high dimensional space. These definitions are based on neighbourhoods between decision regions.
Here, a labelled sample is denoted as $(\mathbf{x}, y)$ with $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathcal{X}$ and $y \in \mathcal{Y}$. $\mathcal{X}$ is a real-valued topological space $(\mathcal{X}, \omega)$ with $\mathcal{X} \subseteq \mathbb{R}^n$ and ω as the Euclidean topology. I will use only $\mathcal{X}$ in the following to refer to the topological space. The application of a classifier c divides the space $\mathcal{X}$ into labelled partitions $\{\mathcal{X}_1, \ldots, \mathcal{X}_{|\mathcal{Y}|}\}$, for $|\mathcal{Y}|$ classes and based on the samples given. Each partition represents a decision region $\mathcal{D}_y$ and might be seen as a labelled topological space.

If for any two points $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ there is a continuous map $f : [0,1] \mapsto \mathcal{X}$, such that $f(0) = \mathbf{x}$ and $f(1) = \mathbf{x}'$, $\mathcal{X}$ is called path-connected.
If a topological subspace $\mathcal{D}_y$ is path-connected, the decision region of class y consists of exactly one component and will be called connected. Otherwise it can be expressed as a set of $|cp|$ path-connected components $\mathcal{D}_y = \{cp_1, \ldots, cp_{|cp|}\}$ with cp_i being path-connected for all i. In this case if $|cp| > 1$ the decision region $\mathcal{D}_y$ will be called disconnected.

A classifier c can hence be considered as partitioning the input space in a set of labelled path-connected components $\mathbf{CP}_c$. The neighbourhood of these components defines the relative location of the class labels in the feature space. These neighbourhoods can be expressed in terms of a component graph G_c (Definition 4.1).

Definition 4.1 (Component graph)**.**
The *component graph* G_c of a classifier c is defined as an undirected simple graph $\mathcal{G}(\mathbf{CP}_c, \mathbf{E}_c)$ with the set of vertices given by the components $\mathbf{CP}_c$ and the set of edges $\mathbf{E}_c \subseteq \{(cp_i, cp_j) : (cp_i, cp_j) \in \mathbf{CP}_c \times \mathbf{CP}_c, i < j\}$ given by the neighbourhood

$$(cp, cp') \in \mathbf{E} \iff \left[\delta(cp) \cap \delta(cp') \setminus \bigcup_{cp'' \in \mathbf{CP}_c \setminus \{cp, cp'\}} \delta(cp'') \right] \neq \emptyset. \tag{43}$$

$\delta(cp)$ is the set of adherent points of cp.

An illustration for constructing a component graph from a set of samples is given in Figure 30.
The graph consists of several paths that can be analysed based on their characteristics. If each label is represented at least once in a path it is called class complete (Definition 4.2). If additionally the order of the labels within the path represents the assumed order on the labels, it is called ordinal (Definition 4.3).

Definition 4.2 (Class complete path)**.**
A path $(cp_1, \ldots, cp_i, \ldots, cp_p)$ in G_c is *class complete* (for $\mathcal{Y}$) if $\forall y \in \mathcal{Y}$, $\exists i : cp_i \subseteq \mathcal{D}_y$.

Definition 4.3 (Ordinal path)**.**
A path $(cp_1, \ldots, cp_i, \ldots, cp_{|\mathcal{Y}|})$ in G_c is *ordinal* (for $y_{(1)} \prec \ldots \prec y_{(|\mathcal{Y}|)}$) if $\forall i$: $cp_i \subseteq \mathcal{D}_{y_{(i)}}$.

To evaluate the type of ordinality the following four criteria can be used. These criteria are ordered according to their strictness. There might be more criteria, like a criterion between criterion II and III, which claims that exactly two classes are represented by connected decision regions. A 4-class example scheme is given in Figure 31.

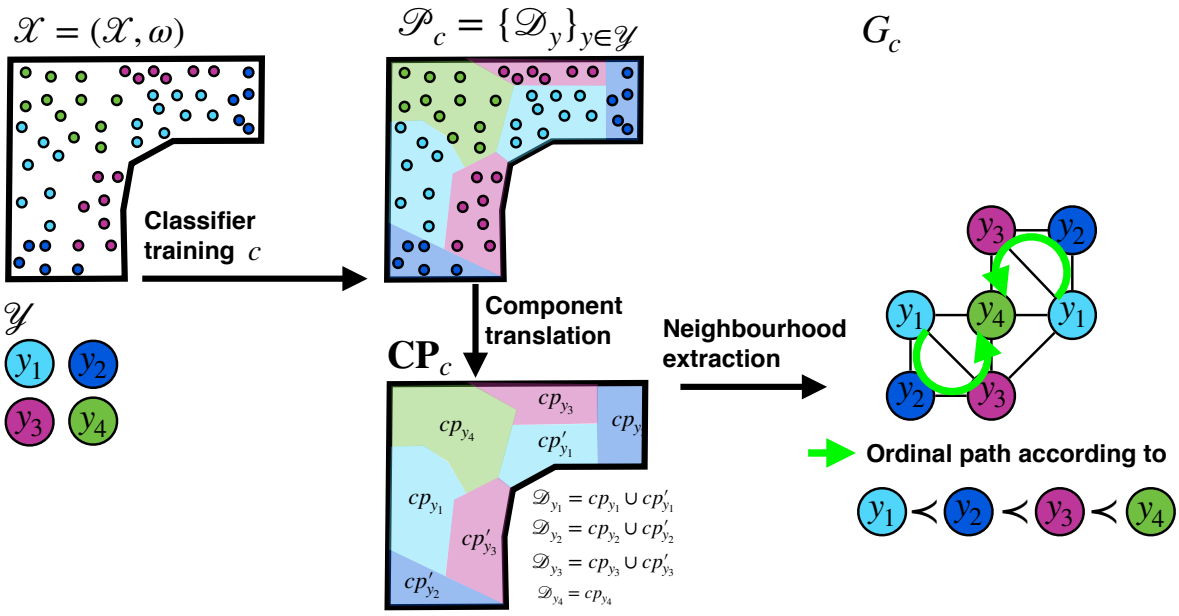


Figure 30: Construction scheme of a component graph using a 4-class example. Each sample is depicted by a small dot and the colour represents its label. Large circles and their colour represent the class labels in the set $\mathcal{Y}$ and the representation as nodes in the component graph. The training of a classifier divides the feature space $\mathcal{X}$, considered as real-valued topological space $(\mathcal{X}, \omega)$ with $\mathcal{X} \subseteq \mathbb{R}^n$ and ω being the Euclidean topology, into a set of decision regions $\mathcal{P}_c = \{\mathcal{D}_y\}_{y \in \mathcal{Y}}$ (Classifier training). Each decision region consists of a set of path-connected components. The set of all components **CP** comprises the whole feature space $\mathcal{X}$ (Component translation). Based on the neighbourhood of these components a labeled component graph g_c for classifier c can be defined (Neighbourhood extraction). This graph might contain ordinal paths according to a predefined order. Here two ordinal paths according to $y_1 \prec y_2 \prec y_3 \prec y_4$ are marked in green.

Criterion I: For each $i \in \{1, \ldots, |\mathcal{Y}|\}$ and for each component $cp_i \in \mathbf{CP}_{y_{(i)}}$ there exists an ordinal path $(cp_1, \ldots, cp_i, \ldots, cp_{|\mathcal{Y}|}) \in G_c$ with $\forall j \neq i$: $cp_j \in \mathbf{CP}_{y_{(j)}}$.

Criterion II: Criterion I is fulfilled and one class is represented by a connected decision region.

Criterion III: Criterion I is fulfilled and all classes are represented by connected decision regions.

Criterion IV: All class complete paths are ordinal paths (according to $y_{(1)} \prec \ldots \prec y_{(|\mathcal{Y}|)}$ or $y_{(|\mathcal{Y}|)} \prec \ldots \prec y_{(1)}$).

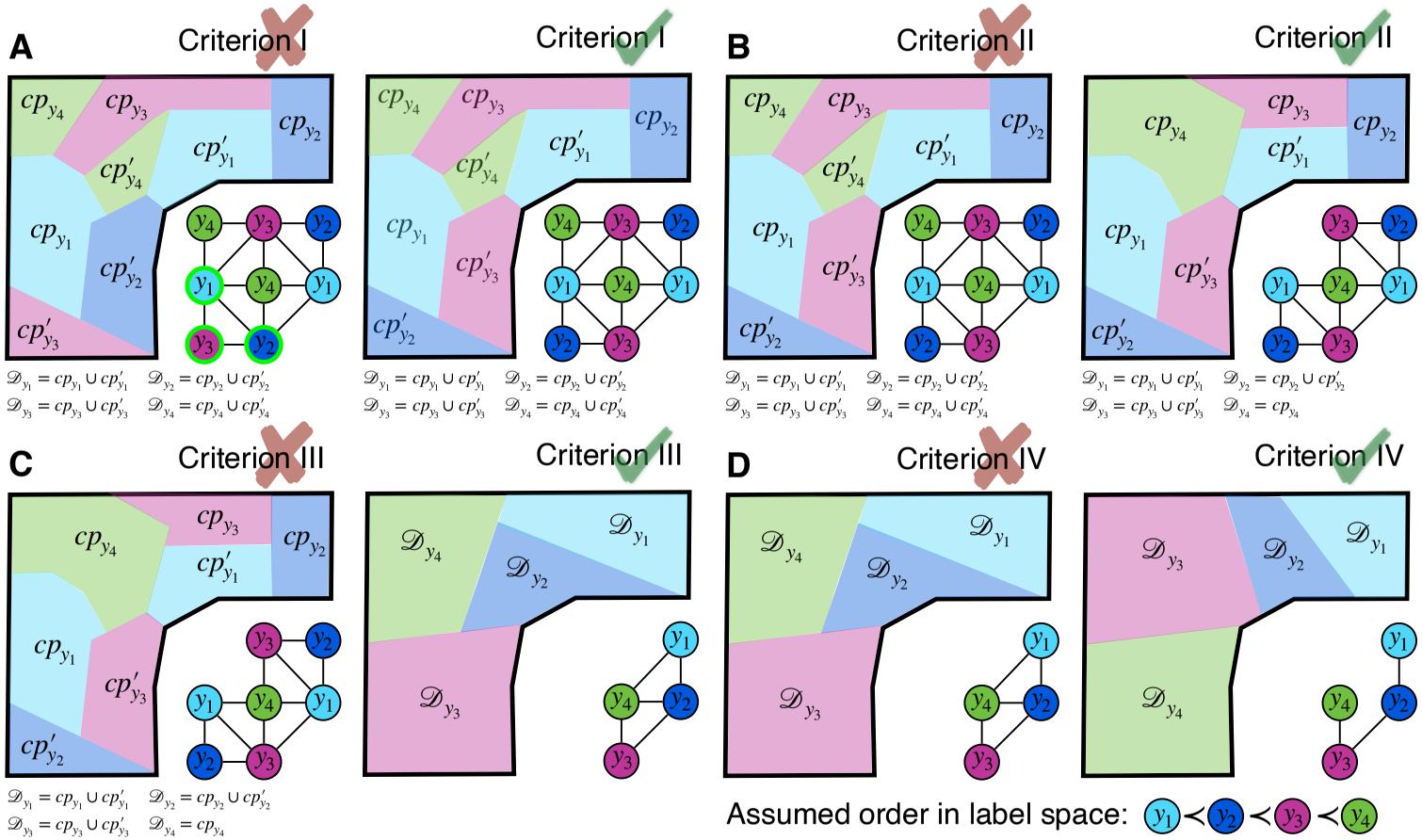


Figure 31: Example scheme of the four ordinality criteria. Each sample is represented by a circle and its label is given by the colour. The order in the label space is given by the order at the bottom right of the scheme (y_1 : lightblue $\prec$ y_2 : darkblue $\prec$ y_3 : purple $\prec$ y_4 : green). The label of each decision region is indicated by the colour. For each criterion two examples are given, the left one shows a representation, that does not fulfil the criterion and the right shows a representation that fulfils the criterion. The corresponding component graph is given next to each feature space representation. For criterion I not fulfilled, samples for which it is not possible to form an ordinal sequence are marked with a green border. See [120] licensed under CC BY-SA 4.0, https://creativecommons.org/licenses/by-sa/4.0/.

4.2 Experimental results

I analysed a range of 4-class artificial datasets constructed by clouds around different centroids. Datasets $d_1, ..., d_6$ show a class label pattern that can be observed by a projection to their first principal component. The patterns of datasets $d_7, ..., d_{10}$ can be observed by a projection to the first and second principal component. The representation in feature space is given by the centroid location. The corresponding patterns, their centroid locations $\mathbf{z}_y$ and the criteria they fulfil by construction are given in Figure 32.

Each sample cloud comprises 50 samples $\mathbf{x}$, that are drawn according to a Gaussian distribution $\mathbf{x} \sim \mathcal{N}(z_y, \sigma \mathbb{I}^n)$ for the 1D patterns with z_y. For the 2D patterns the centroids are $\{z'_y, z''_y\}$ and $\mathbf{x} = (\mathbf{x}'^T, \mathbf{x}''^T)$ with $\mathbf{x}' \sim \mathcal{N}(z'_y, \sigma \mathbb{I}^{n/2})$ and $\mathbf{x}'' \sim \mathcal{N}(z''_y, \sigma \mathbb{I}^{n/2})$. The values of $\{z_y, z'_y, z''_y\}$ are given in Figure 32.

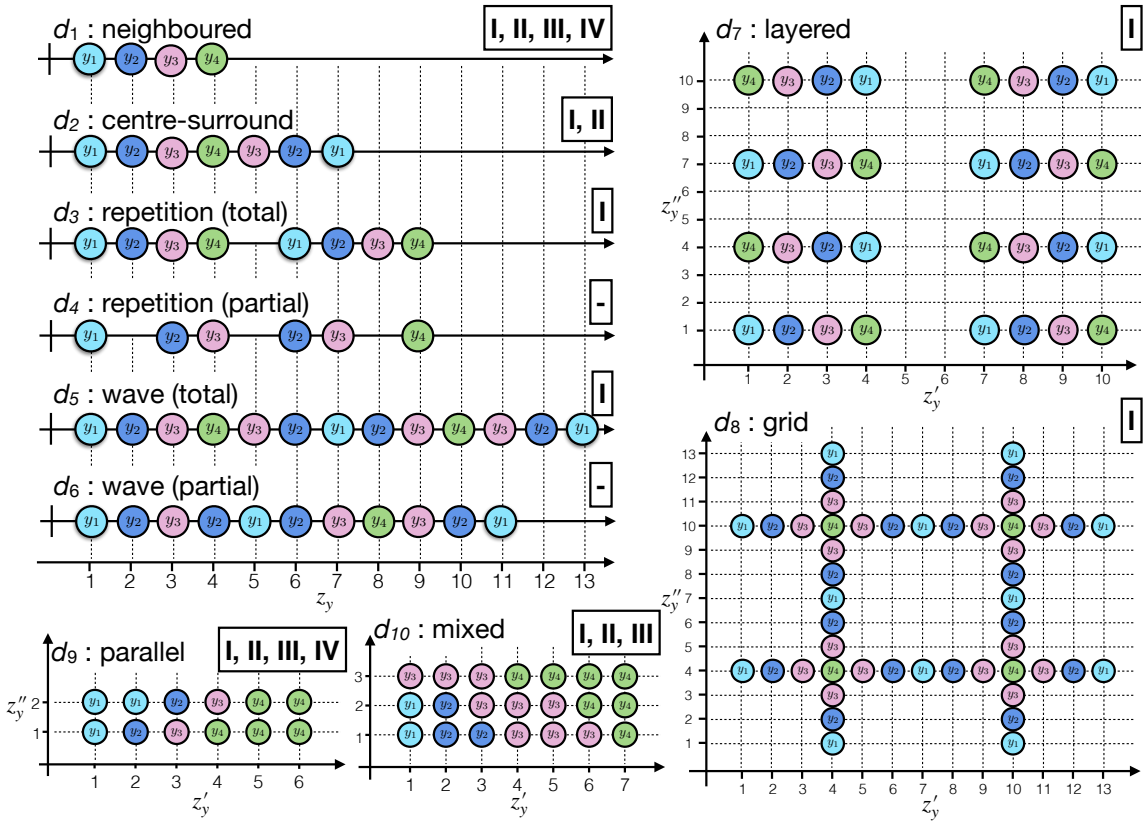


Figure 32: Prototype locations of the diverse sampled artificial ordinal patterns. Each circle represents a cloud of 50 samples. The coordinate systems denote the centroid locations used to sample these patterns. For each dataset the criteria fulfilled are denoted in the black square. This scheme is adapted from [120] licensed under CC BY-SA 4.0, https://creativecommons.org/licenses/by-sa/4.0/.

By scanning the datasets $d_1, ..., d_6$ and $n = 1$ through different standard deviations $\sigma \in \mathbb{R}$, I determined $\sigma = 0.3$ to result in well-separable sample clouds. As dimensions I choose $n = \{2, 10, 50, 100, 500, 1000, 2000, 5000\}$ and for $d_1, ..., d_6$ additionally $n = 1$. I screened these data using 10 x 10 cross-validation (CV) and a linear L2-support vector machine (SVM) and k-nearest neighbour (kNN) (k=3) as base classifier for the Occ and used a threshold of 0.65 on the minimal class-wise sensitivity as performance measure.

The results show a difference in the performance of the 3-nearest neighbour (NN) and the linear SVM (Figure 33). The assumed order in the label space is $y_1 \prec y_2 \prec y_3 \prec y_4$. The 3NN returns for all patterns the order as being reflected in feature space, even for the patterns that show a partial repetition (d_4, d_6) and do not fulfil the weakest ordinality criteria I. The linear SVM as classifier without disconnected decision regions only finds the global ordinal pattern in d_1, hence only those patterns that fulfil at least criteria III.

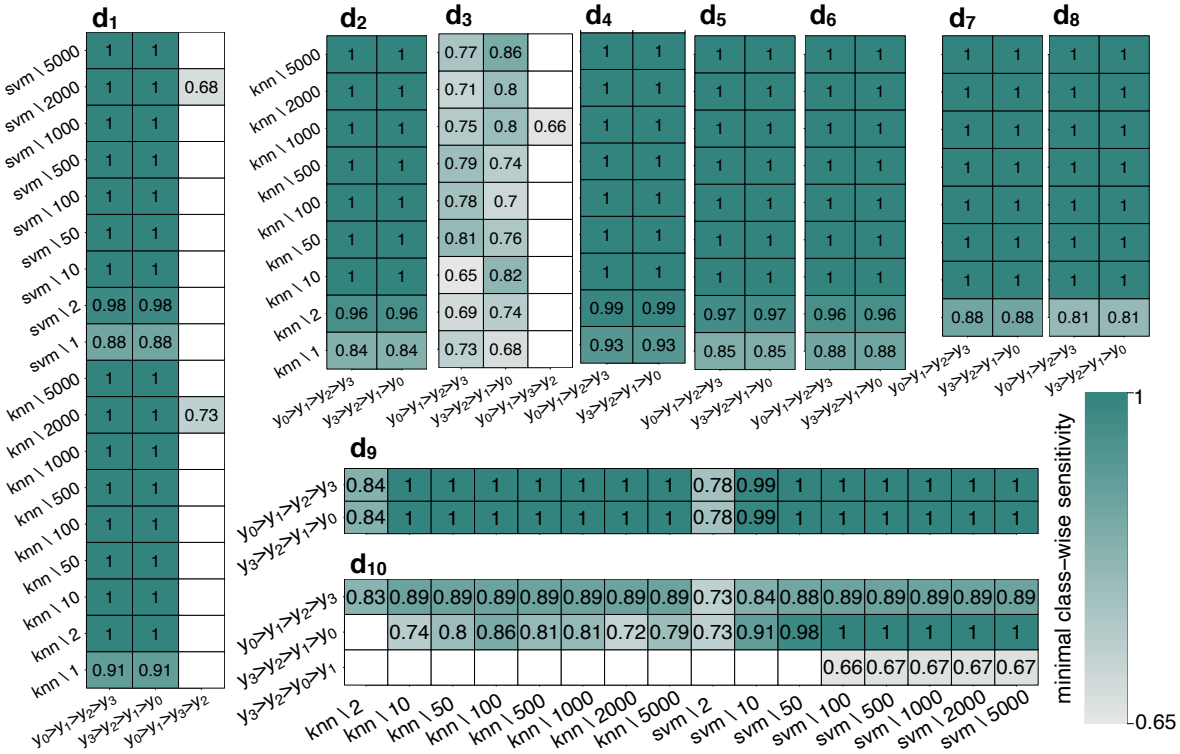


Figure 33: Performance of the returned cascades in diverse artificial presumable ordinal patterns. For each dataset the set of returned cascades per base classifier support vector machine (linear L2-SVM) or k-nearest neighbour (kNN) and each dimension ({1, 2, 10, 50, 100, 500, 1000, 2000, 5000}) passing a threshold of 0.65 is shown. The performance is measured in terms of minimal class-wise sensitivity and depicted as values as well as in a colour coded way.

For the datasets d_9, d_{10} I additionally compared the linear L2-SVM as base classifier with an ensemble of linear L2-SVMs that are restricted to parallel decision boundaries. The results show that the SVM with parallel decision boundaries shows minimal class-wise sensitivities higher than 0.8 for the expected order independent of the dimension for dataset d_9, a dataset that fulfils criteria IV, but not for dataset d_{10}.

4.3 Discussion and conclusion

In order to better understand and interpret the returned relational patterns, the meaning of ordinality in high-dimensional spaces has to be considered. This cannot be understood intuitively and there might be a set of diverse patterns that can be considered as ordinal. One distinction might be drawn by considering global and local ordinal patterns. Within the scope of this thesis I developed a categorisation construct to distinguish between types or ordinality. This construct might be further refined and used to understand the neighbourhood relations of categories, which are found in high-dimensional space. Furthermore, it might serve as framework to structure types of ordinality, as for example the working definition of [25] can be classified as fulfilling at least criterion III.
I proposed categories by using a component graph, and the set of possible paths on this graph. The characteristics of the base classifiers constrain the set of possible graph structures and hence base classifiers might be used to distinguish between different types of ordinality. Experimentally, I showed the different capacity of classifiers that allow for disconnected decision regions and classifiers that do not. The former can find repetitive local ordinal structures with the risk of failing in detecting non-ordinal partial repetitive structures. The later can only detect global ordinal structures.
With this analysis I made a step towards a theoretical analysis of the observed performance differences by using different base classifiers.

5

RESULTS ON REAL DATA SCREENINGS USING THE ORDINAL CLASSIFIER CASCADE

Learning a classifier aims at finding a classifier that best distinguishes between diverse classes. Existing differences can be found by a suitable classifier. This well-performing classifier might be used to differentiate between samples of different classes and to support decision making. The training process might be performed solely based on intrinsic data characteristics given in the feature space. However, also external domain knowledge about the class characteristics might be used to improve classification performance. External domain knowledge refers to knowledge, which is not explicitly given in the feature space. This knowledge can be included into the classification training process by constraining the classifier [38, 65].
Accurate decisions are of high importance in the field of medical diagnosis, as wrong classification might lead to an inappropriate treatment with potential drastic outcome for the patient. The benefit of constraining the classifier relies on the goodness of the additional knowledge and possibly on its reflection in the investigated feature space. The ordinal classifier cascade (Occ) only achieves high performance if the assumed order is reflected in the feature space [119]. Using this, I could show that the power of the Occ is beyond improving classification performance. Indeed, the validation and characterisation in Chapter 3 shows that the Occ can be used to reveal new and potential unknown relations from high-dimensional space.
Screening for orders allows a range of possible applications for biological and medical data, like raising hypotheses about neighbourhoods in subgroups of cancer (Section 5.1), finding the order reflection in developmental stages (Section 5.2) or confirming or rejecting the reflection of orders hypothesised based on histological observations in gene expression data of pre-cancer stages (Section 5.3).

5.1 Revealing orders in static data

Many multi-class datasets comprise various classes belonging to the same research context, as these datasets were collected in order to tackle a specific research question. Based on external domain knowledge one might know, hypothesise or search

for relations between these classes or between a subset of classes. Relations could represent a relation in time, an ordering in similarity or in the context of cancer in severity or tissue disruption. The relations are expressed on a semantic level between the classes or phenotypes and usually the reflection of semantic relations in feature space is only assumed but actually not controlled.
Even if a potential relation between classes might not be known or assumed, it could give a deeper insight into the research context and lead to better understanding. As relations could occur on all classes but also solely on subset of classes, there is the need for an explorative screen for potential relations not only on the whole set of classes but additionally on all subsets. The Occ can be used for this task [122].

5.1.1 Screening results

The screening of two datasets with 7 and 16 classes (d_1, d_2) using the Occ revealed a set of possible candidate cascades for each dataset (Table 5). Graphical representation are given in Figures 34 and 35.

Table 5: Characteristics of the candidate set comprising the longest cascades of static datasets. For each dataset (d_1: Hypothalamic-pituitary-adrenal (HPA)-axis, d_2: Leukaemia), the final threshold on the minimal class-wise sensitivity (t), the length (k), the number of cascades within the set as well as the number of possible cascades of the given length (in brackets) is given. Furthermore, the minimal class-wise sensitivity (smin) and the averaged purity of the alternative (pur_{alt}) and uncovered (pur_{unc}) classes are shown in %. Alternative classes comprise classes that are not part of the current cascade but part of any cascade in the candidate set. Uncovered classes are classes that are not part of any of the proposed cascades. The size of each set is given as alt and unc. See [122] licensed under CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

candidate set	smin	pur_{alt}	pur_{unc}
d_1: **HPA-axis** (t = 0.9, k = 5): 5 (2520)		alt=1	unc=1
$y_7 \prec y_6 \prec y_2 \prec y_4 \prec y_5$	93.1 %	85.6 %	68.3 %
$y_7 \prec y_6 \prec y_2 \prec y_5 \prec y_4$	93.1 %	95.2 %	68.3 %
$y_5 \prec y_4 \prec y_2 \prec y_6 \prec y_7$	92.7 %	87.7 %	89.8 %
$y_5 \prec y_4 \prec y_3 \prec y_6 \prec y_7$	92.3 %	90.8 %	43.3 %
$y_7 \prec y_6 \prec y_3 \prec y_4 \prec y_5$	91.4 %	90.4 %	36.5 %
d_2: **Leukaemia** (t= 0.75, k= 7): 2 ($> 10^{18}$)		alt=6	unc=5
$y_{12} \prec y_2 \prec y_{10} \prec y_{18} \prec y_5 \prec y_9 \prec y_8$	75.8 %	69.3 %	61.1 %
$y_1 \prec y_{17} \prec y_5 \prec y_7 \prec y_{13} \prec y_{15} \prec y_{16}$	75 %	77.0 %	69.8 %

A value of 0.7 was set as minimal requirement on the performance. A candidate set comprised the longest cascades that pass the highest threshold, which was found by increasing the threshold of 0.7 in steps of 0.05. For none of these datasets a relation between all classes was known or assumed beforehand. But there was a relation between three classes of d_1 reported [7].

The explorative screen showed that the candidate sets of both datasets comprised relations between subsets of classes.

For dataset d_1 (hypothalamic-pituitary-adrenal (HPA)-axis) five cascades of length 5 passed a threshold of 0.9, which is less than 0.2 % of all cascades of length 5 (Figure 34). One cascade showed a potential swap in the last two classes ($y_7 \prec y_6 \prec y_2 \prec y_5 \prec y_4$). The other cascades were returned in both directions, indicating a symmetry. These cascades differed only in their centre classes (y_2 versus y_3) but showed a high purity for these alternative classes being assigned towards each other. There is one class y_1, the cortex, which was not included in any cascade and had a purity ranging from 36.5 % till 89.8 % (Figure 34B, Table 5).

The leukaemia dataset (d_2) revealed two cascades of length 7 passing a threshold of 0.75 (Figure 35). Both cascades overlapped at exactly one class y_5, showing a high purity for the alternative and the uncovered classes (Table 5). Within the cascades a grouping of myeloid leukaemia (ML) and lymphocytic leukaemia (LL) classes could be observed. This grouping was additionally seen within the assignment of the external classes for the second cascade (Figure 35B).

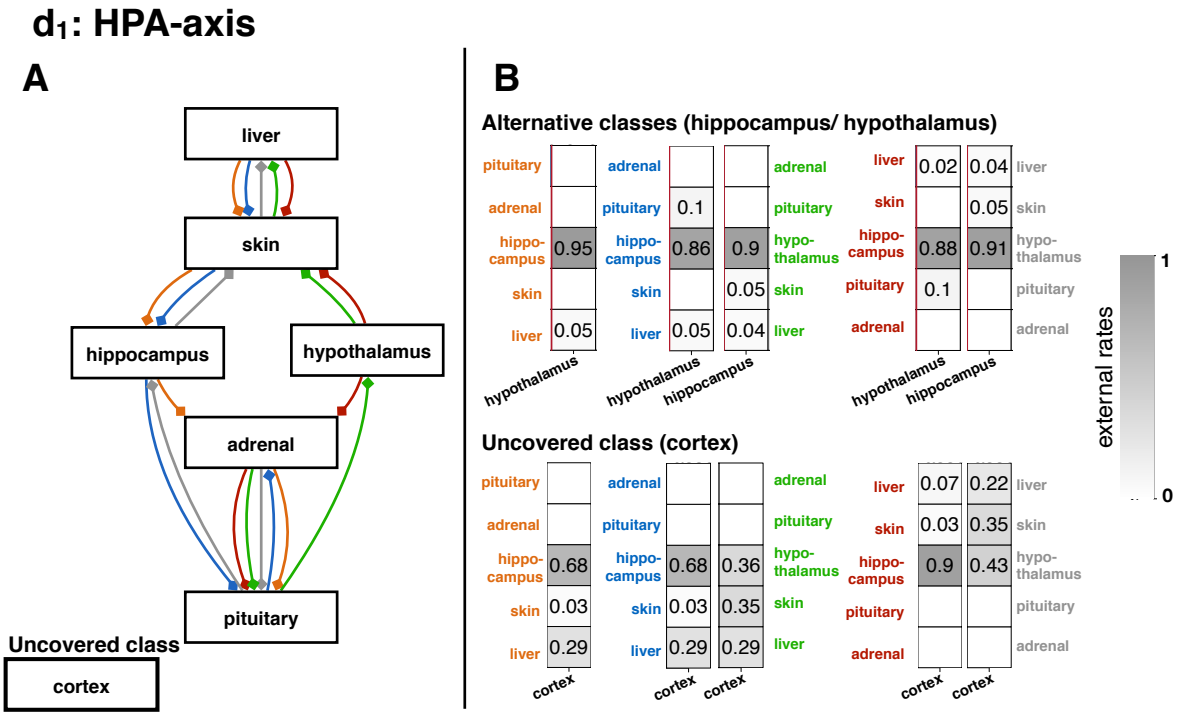


Figure 34: Cascade results on the hypothalamic-pituitary-adrenal (HPA)-axis dataset (d_1). For each cascade A: a graphical representation of the proposed candidate set as well as B: the confusion tables for all classes that are not part of the cascade (alternative and uncovered classes) are shown. The proposed set comprises five cascades with a length of 5. The order within the cascades is shown in the graph and the external class is given at the bottom. The cascades are colour coded. For each cascade one heatmap for the alternative and one for the uncovered class is given with the cascade classes given in the rows and colour coded in accordance with the graph. The columns comprise the alternative and uncovered classes, respectively and the values correspond to the external rates indicating the amount of samples of the class given in the column that is assigned to the class given in the row. See [122] licensed under CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

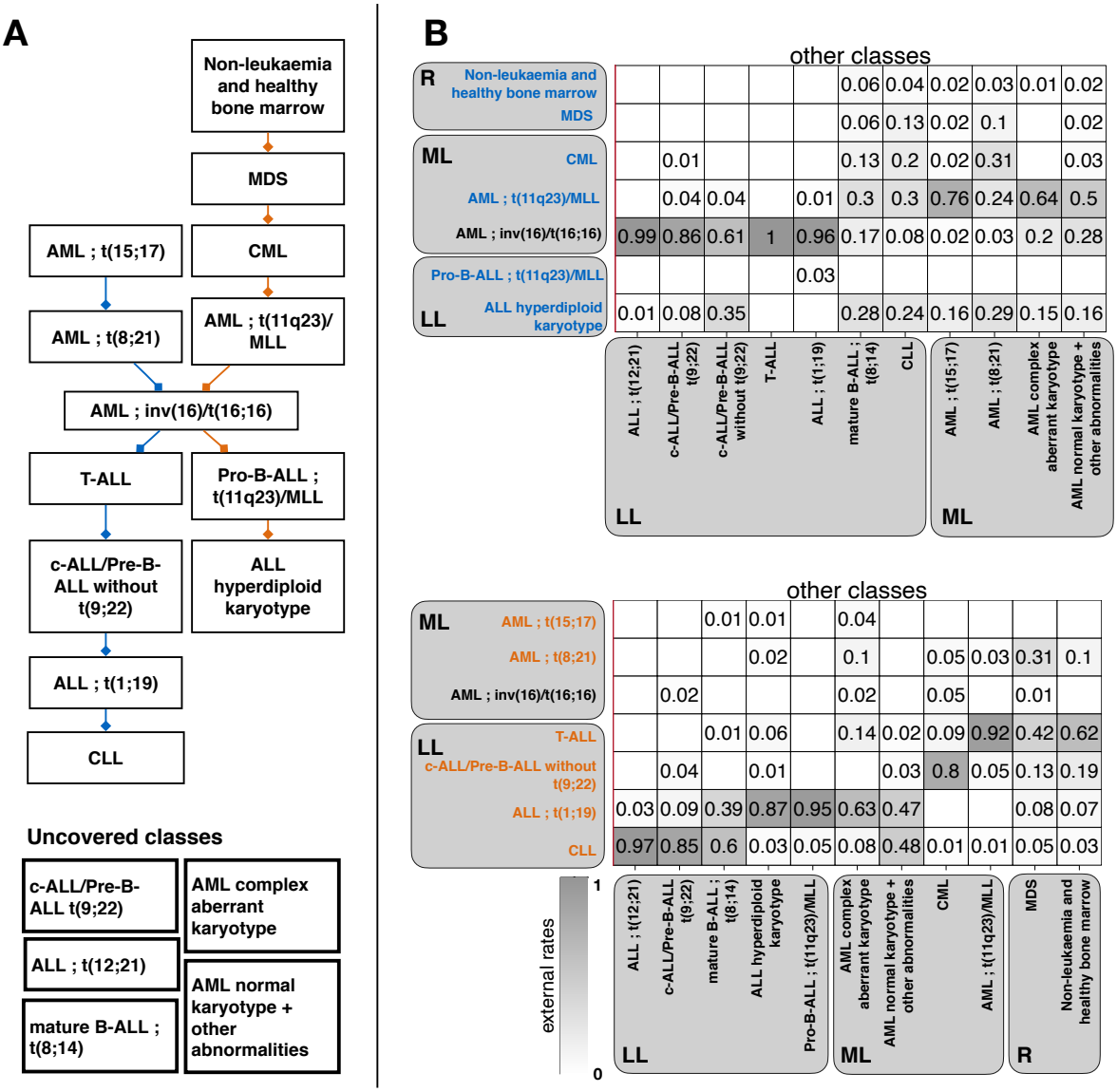


Figure 35: Cascade results on the leukaemia dataset (d_2). For each cascade A: a graphical representation of the proposed candidate set and B: the confusion tables for all classes that are not part of the cascade (external classes) are shown. The proposed set comprises two cascades with a length of 7, which overlap in one class. The order within the cascades is shown in the graph and the uncovered classes are given at the bottom. The cascades are colour coded. For each cascade one heatmap is given with the cascade classes given in the rows and colour coded in accordance with the graph. The black coded class belongs to both returned cascades. The classes are grouped by myeloid leukaemia (ML), lymphocytic leukaemia (LL) and rest (R). The columns comprise all remaining classes (alternative and uncovered classes) also grouped in ML and LL and the values correspond to the external rates indicating the amount of samples of the class given in the column that is assigned to the class given in the row. See [122] licensed under CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

5.2 Cascades in developmental stages of model organisms

In addition to exhaustive screens on datasets with static classes, screenings on datasets that represent developmental stages of model organisms were performed. Hereby, it was expected to reveal the time ordering between all classes from the feature space and thereby showing its reflection. The samples of one dataset were additionally to categorising in stages, labelled by the time points taken.
All total cascades passing an initial bound of 0.5 were reported. As this bound is an upper bound on the overall performance (Theorem 1), it is used to reduce the number of cascades that have to be evaluated.
The number of cascades that have been rejected due to the thresholding as well as the final cascades and their minimal class-wise sensitivity are given in Table 6. The bound of 0.5 is used for preselection and the desired performance might be higher, like for example 0.7.

Table 6: Set of returned cascades for datasets of developmental stages of model organisms (d_3, d_4, d_5, d_6). For each dataset all permutations of orders were screened and all cascades passing a screening bound on the sensitivity of $t \geqslant 0.5$ are given. This bound defines an upper threshold for the performance in terms of minimal class-wise sensitivity. The number of cascades that were rejected due to this bound is given in brackets. The remaining cascades were evaluated and the minimal class-wise sensitivity (smin) is denoted. See [123] licensed under CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

ordinal cascades	bound	smin
d_3: ***Drosophila melanogaster*** (21 total cascades rejected)		
$\mathrm{embryo} \prec \mathrm{larva} \prec \mathrm{pupa} \prec \mathrm{adult}$	94.4 %	89.4 %
$\mathrm{adult} \prec \mathrm{pupa} \prec \mathrm{larva} \prec \mathrm{embryo}$	72.3 %	72.3 %
$\mathrm{adult} \prec \mathrm{pupa} \prec \mathrm{embryo} \prec \mathrm{larva}$	71.0 %	71.0 %
d_4: ***Danio rerio*** (117 total cascades rejected)		
$\mathrm{adult}_2 \prec \mathrm{adult}_1 \prec \mathrm{embryo}_3 \prec \mathrm{embryo}_2 \prec \mathrm{embryo}_1$	85.0 %	85.0 %
$\mathrm{adult}_2 \prec \mathrm{adult}_1 \prec \mathrm{embryo}_3 \prec \mathrm{embryo}_1 \prec \mathrm{embryo}_2$	85.0 %	85.0 %
$\mathrm{adult}_2 \prec \mathrm{adult}_1 \prec \mathrm{embryo}_1 \prec \mathrm{embryo}_2 \prec \mathrm{embryo}_3$	54.7 %	54.7 %
d_5: ***Caenorhabditis elegans* (*C. elegans*), stages** (116 total cascades rejected)		
$\mathrm{stage}_1 \prec \mathrm{stage}_2 \prec \mathrm{stage}_3 \prec \mathrm{stage}_4 \prec \mathrm{stage}_5$	91.7 %	91.7 %
$\mathrm{stage}_5 \prec \mathrm{stage}_4 \prec \mathrm{stage}_3 \prec \mathrm{stage}_2 \prec \mathrm{stage}_1$	75.0 %	66.7 %
$\mathrm{stage}_1 \prec \mathrm{stage}_2 \prec \mathrm{stage}_3 \prec \mathrm{stage}_5 \prec \mathrm{stage}_4$	58.6 %	58.6 %
$\mathrm{stage}_5 \prec \mathrm{stage}_4 \prec \mathrm{stage}_3 \prec \mathrm{stage}_1 \prec \mathrm{stage}_2$	50.0 %	50.0 %
d_6: ***Caenorhabditis elegans* (*C. elegans*), time** ($10! - 1$ total cascades rejected)		
$t_1 \prec t_2 \prec t_3 \prec t_4 \prec t_5 \prec t_6 \prec t_7 \prec t_8 \prec t_9 \prec t_{10}$	66.7 %	66.7 %

Further refining the threshold on the final performance to 0.7 it can be observed that for the first three datasets (d_3, d_4, d_5) the expected order of staging was included within the set of cascades and more not expected orders were sorted out. Other cascades that were still included were either the reverse direction or they showed a swap within the last two classes. This swap is, as it was indicated by the screening on the artificial data (Section 3.3), most likely due to technical reasons as there are less constraints imposed the later a class is included into the cascade.
The only dataset that would return an empty set of total cascades imposing a threshold of 0.7 is the time staging of *C. elegans* (d_6). Nevertheless, for this dataset two cascades of length 9 would be returned ($t_1 \prec t_3 \prec t_4 \prec t_5 \prec t_6 \prec t_7 \prec t_8 \prec t_9 \prec t_{10}$, $t_1 \prec t_2 \prec t_4 \prec t_5 \prec t_6 \prec t_7 \prec t_8 \prec t_9 \prec t_{10}$), both with a minimal class-wise sensitivity of 80.0 % and a purity of 86.7 % and 83.3 %, respectively. One cascade does not include the samples taken at minute 23 (y_1) and the other does not include the ones taken at minute 41 (y_2). The confusion tables (Appendix Figure A5) show that these external classes were assigned to their assumed neighbours and the general time ordering could still be found.
Furthermore, it can be seen that the optimal threshold defining the possible candidate set might be individually refined for each dataset. For example a threshold of $t > 85.0$ % imposed on all datasets results in only the expected cascade for d_3 (*Drosophila melanogaster*) and might be considered as optimal threshold for this dataset but would exclude all total cascades for d_4 (*Danio rerio*).

5.3 Cascades in datasets with (pre-)cancer stages

Many cancer types are proposed to develop through a sequence of precursor lesions [228]. The lesions are mostly staged based on histological observations and ordered according to the grade of tissue abnormalities rather than the point in time they occur. The reason is that the follow up of a lesion would require frequent biopsies and the not-extraction of a possible cancerous tissue within a patient. Both are infeasible for the patient and highly unethical. As, however, the phenotype is related to the underlying molecular mechanism, one assumes that the relationship defined on the histological observations is reflected in the gene expression profile.

5.3.1 Screening settings

I screened through stages of three different carcinoma types with different granularity of staging in order to examine whether the ordering assumption holds for the given measurements. Hereby, in addition to using the linear L2-support vector machine (SVM) as base classifier, I performed screenings using the linear L1-SVM to take advantage about its intrinsic feature selection and gain insight about poten-

tial gene sets. Nilsson et al. [160] showed that the optimisation solution for a linear SVM in high dimensions is independent of the chosen hyperparameter. Using linear L1-SVM the hyperparameter might influence the result. I therefore performed a base classifier hyperparameter tuning (cost parameter C) on accuracy as well as sensitivity of the first class by using the *TunePareto* R-package [158]. I call these settings accuracy tuned ordinal classifier cascade (aOcc) and sensitivity tuned ordinal classifier cascade (sOcc), respectively. This results in type-homogenous but not hyperparameter homogenous cascades. I consider cascades as hyperparameter homogenous if they are made up of base classifiers that have the same hyperparameter value. To result in this type of cascades, I furthermore optimised the hyperparameter on the cascades rather than the binary classifier level. I screened through a range of cost parameters (0.1, 0.2, ..., 1, 2, ..., 10, 15, ..., 100) that were set the same for all binary classifiers and chose the cost parameter that resulted in the highest number of pairwise sensitivities exceeding a threshold of 0.7. I call this setting homogenous ordinal classifier cascade (hOcc).
Overall, I used two types of base classifier (linear L1-SVM, linear L2-SVM) and three strategies of hyperparameter tunings (accuracy tuned, sensitivity tuned, homogenous). My screening resulted in a set of cascades. For refining the final candidate set I furthermore analysed two strategies: I selected either the longest and best once with minimal class-wise sensitivity smin_{blc} and a set made up of cascades in a range of $\text{smin} \geqslant \text{smin}_{\text{blc}} - 0.05$, or the Pareto set in terms of length and minimal class-wise sensitivity. An overview of the used settings is given in Figure 36.

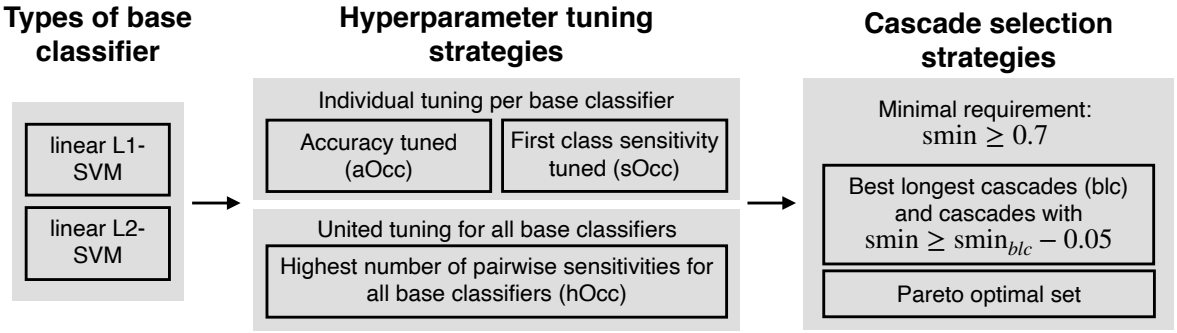


Figure 36: Overview of the analysed screening settings. The following settings were tested: Two different kinds of base classifier (linear L1-support vector machine (SVM), linear L2-SVM); Three types of hyperparameter tuning strategies, of which two perform an individual tuning per base classifier and might lead to a cascade consisting of base classifiers with differing hyperparameters and one performs a tuning on all base classifiers; Two cascade selection strategies. smin is the minimal class-wise sensitivity.

5.3.2 Parameter robustness of homogenous ordinal classifier cascade

In order to gain insight into the influence of the cost parameter in the homogenous ordinal classifier cascade (hOcc) setting on the selection of candidate cascades, I compared the frequency of all cascades that were selected for more than 10 of the 37 different cost parameters (0.1,0.2, ... 1, 2, ..., 10, 15, ..., 100). I analysed the robustness of the cascades for the selected cost parameter with respect to the other hyperparameter values. An overall robust result for linear L2-SVM was shown. Independent of the cost parameter the same set of longest cascades passing $t = 0.7$ and the same Pareto set were returned for all cost parameters. This was not the case for linear L1-SVM (Table 7). However, the results of the selected cost parameter were among the most often selected once and only $y_0 \prec y_1 \prec y_3 \prec y_2$ for the colon carcinoma dataset (d_7) was found for less than 27 % (8/37) of the cost parameters in the Pareto set, indicating a less robust solution.

5.3.3 Screening results

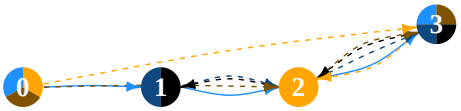
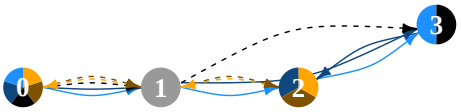
The screening results of the linear L2-SVM and the hOcc using linear L1-SVM selecting the Pareto set is shown in Table 8. The results of the remaining optimisation methods and the set of longest cascades is given in Appendix Tables A1 and A2. The robustness of the linear L2-SVM concerning the cost parameter led to the same results, independent of the kind of hyperparameter tuning (accuracy, first class sensitivity, homogenous). Furthermore, it was robust against the selected cost parameter in the homogenous (hOcc) setting (Section 5.3.2).

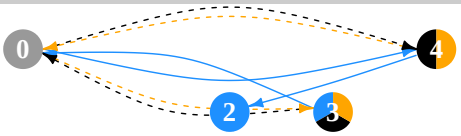
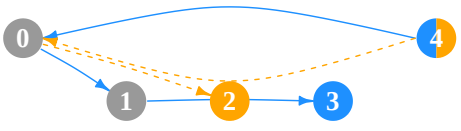
Table 7: Robustness of returned cascades screening the pre-cancer datasets against the cost parameter using linear L1-support vector machine (SVM) homogenous ordinal classifier cascade (hOcc). The cascades that were selected for more than 10 of the 37 (0.1, 0.2, ... 1, 2, ..., 10, 15, ..., 100) different cost parameters are reported for two different selection methods. Once the Pareto optimal solutions are shown and once all cascades passing a minimal class-wise sensitivity of 0.7. Furthermore, their absolute frequency N° is denoted. The cascades that were included in the final set for the selected cost parameter (d_7: Colon carcinoma: $C = 2$, d_8: Pancreatic adenocarcinoma (PDAC), Microarray: $C = 0.9$, d_9: Hepatocellular carcinoma (HCC): $C = 6$) are marked in boldface. The cost parameter was selected based on the the highest number of pairwise sensitivities of all binary classifiers exceeding a threshold of 0.7.

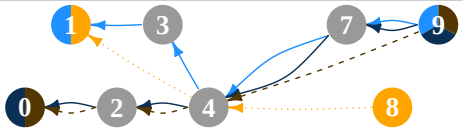
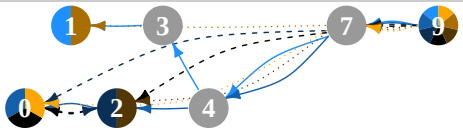
Pareto optimal set					
d_7: Colon carcinoma	N°	d_8: PDAC, Microarray	N°	d_9: HCC	N°
$\mathbf{y_0 \prec y_1 \prec y_2 \prec y_3}$	26	$\mathbf{y_4 \prec y_0 \prec y_1 \prec y_3}$	17	$\mathbf{y_9 \prec y_7 \prec y_0 \prec y_2}$	25
$y_2 \prec y_1 \prec y_0$	22	$y_3 \prec y_0 \prec y_1 \prec y_4$	17	$\mathbf{y_9 \prec y_7 \prec y_1}$	23
$\mathbf{y_0 \prec y_1 \prec y_3}$	22	$y_4 \prec y_0 \prec y_3$	16	$\mathbf{y_9 \prec y_7 \prec y_4 \prec y_3 \prec y_1}$	18
$y_0 \prec y_1 \prec y_2$	22	$y_3 \prec y_0 \prec y_4$	16	$\mathbf{y_9 \prec y_7 \prec y_4 \prec y_2 \prec y_0}$	17
$y_3 \prec y_2 \prec y_1$	10	$\mathbf{y_4 \prec y_0 \prec y_2}$	15	$\mathbf{y_9 \prec y_7 \prec y_2 \prec y_0}$	17
$y_1 \prec y_2 \prec y_3$	10	$y_2 \prec y_1 \prec y_4$	13	$\mathbf{y_9 \prec y_7 \prec y_0}$	15
$y_0 \prec y_3 \prec y_2$	10			$\mathbf{y_9 \prec y_7 \prec y_2}$	11
$y_0 \prec y_2 \prec y_3$	10				

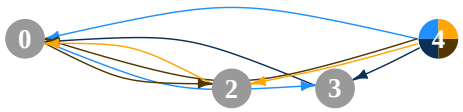
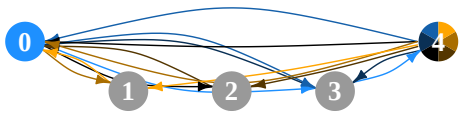
Set of longest cascades $t \geqslant 0.7$					
d_7: Colon carcinoma	N°	d_8: PDAC, Microarray	N°	d_9: HCC	N°
$\mathbf{y_0 \prec y_1 \prec y_2 \prec y_3}$	27	$y_2 \prec y_0 \prec y_1 \prec y_4$	22	$\mathbf{y_9 \prec y_7 \prec y_4 \prec y_2 \prec y_0}$	30
$\mathbf{y_0 \prec y_1 \prec y_3 \prec y_2}$	20	$\mathbf{y_4 \prec y_0 \prec y_1 \prec y_3}$	19	$\mathbf{y_9 \prec y_7 \prec y_4 \prec y_3 \prec y_1}$	24
$y_0 \prec y_1 \prec y_2$	10	$y_3 \prec y_0 \prec y_1 \prec y_4$	17	$y_9 \prec y_7 \prec y_3 \prec y_2 \prec y_4$	19
		$y_4 \prec y_0 \prec y_2 \prec y_3$	14	$y_9 \prec y_6 \prec y_4 \prec y_2 \prec y_0$	14
		$y_4 \prec y_1 \prec y_2 \prec y_3$	13	$y_9 \prec y_7 \prec y_4 \prec y_2 \prec y_1$	12
		$y_4 \prec y_1 \prec y_3 \prec y_2$	10	$y_9 \prec y_7 \prec y_3 \prec y_2 \prec y_0$	12
				$y_8 \prec y_6 \prec y_3 \prec y_2 \prec y_4$	10

Table 8: Screening results on the pre-cancer datasets selecting the Pareto optimal solutions. The results for linear L2-support vector machine (SVM) and linear L1-SVM with a homogenous cost parameter (C) are given. The linear L2-SVM resulted in the same set independent of the optimisation method. The results of further optimisation methods for linear L1-SVM are given in the Appendix Table A1. For each set the performance is denoted by the minimal class-wise sensitivity (smin) in [%]. The graphical illustration shows an arrangement of all cascades found. Each cascade is represented by one colour. The different lengths are given by the line type. Nodes that indicate the first position within a cascade are coloured based on the corresponding cascade, of that this node represents the first position. Nodes that are not used at the first position are coloured grey and cascades that are included for both base classifiers are marked boldface.

d_7: Colon carcinoma			
cascades	smin [%]		graph
L2-SVM, all methods			
$\mathbf{y_0 \prec y_1 \prec y_2 \prec y_3}$	= 94.1	—	
$y_1 \prec y_2 \prec y_3$, $y_3 \prec y_2 \prec y_1$, $y_0 \prec y_2 \prec y_3$, $y_0 \prec y_3 \prec y_2$	= 100	- - -	
L1-SVM, homogenous, C=2			
$\mathbf{y_0 \prec y_1 \prec y_2 \prec y_3}$, $y_0 \prec y_1 \prec y_3 \prec y_2$	= 72.7	—	
$y_0 \prec y_1 \prec y_3$, $y_0 \prec y_1 \prec y_2$, $y_2 \prec y_1 \prec y_0$	= 94.1	- - -	

d_8: Pancreatic adenocarcinoma (PDAC), Microarray			
cascades	smin [%]		graph
L2-SVM, all methods			
$y_3 \prec y_0 \prec y_4 \prec y_2$	= 75	—	
$y_3 \prec y_0 \prec y_4$, $y_4 \prec y_0 \prec y_3$	= 90	- - -	
L1-SVM, homogenous, C=0.9			
$y_4 \prec y_0 \prec y_1 \prec y_3$	= 83.3	—	
$y_4 \prec y_0 \prec y_2$	= 100	- - -	

d_9: Hepatocellular carcinoma (HCC)			
cascades	smin [%]		graph
L2-SVM, all methods			
$y_9 \prec y_7 \prec y_4 \prec y_3 \prec y_1$, $y_9 \prec y_7 \prec y_4 \prec y_2 \prec y_0$	= 77.8	—	
$y_9 \prec y_4 \prec y_2 \prec y_0$	= 90.9	- - -	
$y_8 \prec y_4 \prec y_1$	= 100	·····	
L1-SVM, homogenous, C=6			
$y_9 \prec y_7 \prec y_4 \prec y_3 \prec y_1$, $y_9 \prec y_7 \prec y_4 \prec y_2 \prec y_0$	= 77.8	—	
$y_9 \prec y_7 \prec y_0 \prec y_2$, $y_9 \prec y_7 \prec y_2 \prec y_0$	= 92.3	- - -	
$y_9 \prec y_7 \prec y_2$, $y_9 \prec y_7 \prec y_1$, $y_9 \prec y_7 \prec y_0$	= 100	·····	

d_{10}: PDAC, RNA-Sequencing			
cascades	smin [%]		graph
L2-SVM, all methods			
$y_4 \prec y_0 \prec y_3$, $y_4 \prec y_3 \prec y_0$, $y_4 \prec y_0 \prec y_2$, $y_4 \prec y_2 \prec y_0$	= 100	—	
L1-SVM, homogenous, C=0.1			
$y_0 \prec y_3 \prec y_4$, $y_4 \prec y_0 \prec y_3$, $y_4 \prec y_3 \prec y_0$, $y_4 \prec y_0 \prec y_2$, $y_4 \prec y_2 \prec y_0$, $y_4 \prec y_0 \prec y_1$, $y_4 \prec y_1 \prec y_0$	= 100	—	

For the colon carcinoma (d_7), there was one cascade always included in the set of selected cascades (Table 8), independent of the base classifier, the hyperparamter tuning method or the selection procedure. This was a total cascade of type $y_0 \prec y_1 \prec y_2 \prec y_3$. The second cascade of length 4 showed a swap in the last two classes. This swap was shown as less robust against the cost parameter (Section 5.3.2). Additionally, the artificial data (Section 3.3, Figure 25) showed that this swap often occurs in well-separable classes of a total order. The shorter cascades were either direct or inverse parts (with and without swap in the last classes) of the longer cascades.

In contrast to this, the screening results of the PDAC Microarray dataset (d_8) did not show one clear pattern (Table 8). No total order of length 5 was returned. Most often included in the selected set was a cascade of type $y_4 \prec y_0 \prec y_1 \prec y_3$.

Furthermore, cascades of type $y_3 \prec y_0 \prec y_1 \prec y_4$ and $y_3 \prec y_0 \prec y_4 \prec y_2$ were observed. Within the Pareto set selection the shorter cascades mainly included the external classes, that have not been included in the longest cascades, but the general order of the classes observed in the longer cascade remained. This was not the case for the selection of the longest cascades and their sensitivity range, as the linear L2-SVM screening results in $y_3 \prec y_0 \prec y_4 \prec y_2$ and $y_2 \prec y_3 \prec y_4$ and $y_4 \prec y_2 \prec y_3$. Nevertheless, a general robust trend was observed. Firstly, y_1 and y_2 were rarely used in the same cascade. Secondly, if y_0 was included in the cascade than y_4 and y_3 were always on opposite sites. Fixing y_0 as a root node, the screening results indicated a tree-like structure with two branches.

No total cascade comprising all classes was returned for d_9, the HCC dataset (Table 8). Classes that were often not included in neither of the selected cascades are y_5 and y_6. These classes never occurred in the Pareto set. Classes that were always within cascades of the selected set are y_0, y_1, y_2, y_7 and y_9. It could also be observed that some classes never occured together in a cascade, like y_8 and y_9. The general trend that the classes were ordered either incrementally or decrementally mostly stayed robust against the optimisation procedure, with a few exceptions for not hyperparameter-homogenous optimisation techniques (aOcc, sOcc). These techniques returned a cascade with y_0 placed in the middle (Appendix Tables A1, A2). The decremental way with starting at y_9 was most often observed. Also within this analysis it could be observed that the general order of the classes within the longest cascades was not contradicted by shorter subcascades.

For the ribonucleic acid (RNA)-Sequencing PDAC dataset (d_{10}) only cascades of length 3 were returned. The candidate set comprised cascades of class subgroups that always included class y_0. This class contains samples that are merged to the remaining four classes and the shown relations might not be cancer development related. Excluding this class from the analysis a cascade of type $y_2 \prec y_3 \prec y_4$ with sensitivities of (40.0 %, 77.8 % and 90.9 %) was returned, which might give a hint about a reflection of the progression but is due to its very low performance and the shown result that cascades of length 3 are prone to be false positives (Section 3, Table 4) possibly unreliable.
Due to the merging of experiments, the short cascades and a general very small sample size I will exclude this dataset (d_{10}) from the following analysis but still give its characteristics in the Appendix.

5.3.4 Analysis of possible candidate cascades

To further refine the possible candidate set, I concentrated on the best performing longest cascades exceeding a threshold on the minimal class-wise sensitivity of 0.7, or in other words on the longest cascades in the Pareto optimal set. Except for PDAC, Microarray dataset (d_8) this set returned in the linear L1-SVM hOcc setting was also returned using linear L2-SVM as base classifier. This observation in combination with the attempt to analyse potential gene sets and the effort to impose the same allowance for misclassification within each step of the Occ led to focusing on the linear L1-SVM hOcc setting and the longest cascades of the Pareto set in the following.

5.3.4.1 Pairwise separability and analysis of external classes

A prerequisite for a cascade is the pairwise separability of the classes. It could be observed that for the colon carcinoma (d_7) and PDAC, Microarray (d_8) dataset all classes were separable with class-wise sensitivities exceeding the threshold of 0.7 (Figure 37). This was not the case for the fine granular HCC dataset (d_9). Especially, y_6 was not well-separable from many other classes. But also y_5 and y_8 were less well-separable. These classes were also the classes that are less often included into a cascade (Appendix Tables A1, A2).

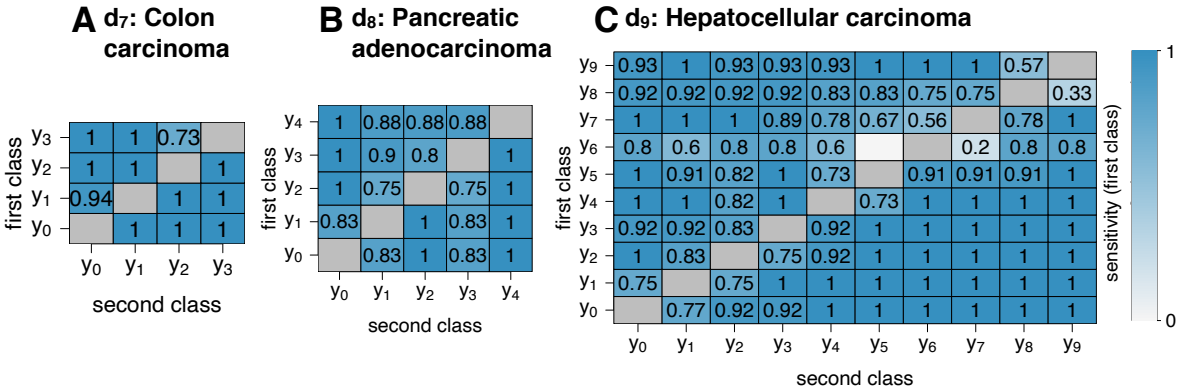


Figure 37: Base classifier performance on the pre-cancer datasets for linear L1-support vector machine (SVM) and a homogenous cost parameter (homogenous ordinal classifier cascade (hOcc)). The heatmaps show the sensitivity of the first class for each binary classifier and the cost parameter that showed the highest number of first class sensitivities above a threshold of $t = 0.7$. A: Results on the colon carcinoma dataset (d_7) with a cost parameter of $C = 2$. B: Results on the pancreatic adenocarcinoma (PDAC) Microarray dataset (d_8) with a cost parameter of $C = 0.9$. C: Results on the hepatocellular carcinoma (HCC) dataset (d_9) with a cost parameter of $C = 6$.

The location of the external classes with respect to the cascade can be expressed in terms of external rates. In the case of d_8 (PDAC, Microarray) y_2 was most often assigned to y_3 but second most often to y_0 rather than y_1. For d_9 (HCC) and the cascade $y_9 \prec y_7 \prec y_4 \prec y_3 \prec y_1$, y_2 spread across y_3 and y_4 rather than y_1 and y_3. But the confusion tables show that in the cascades returned by d_8, as well as in d_9 the samples of the external classes were mainly assigned to their expected neighbouring classes (Figure 38).

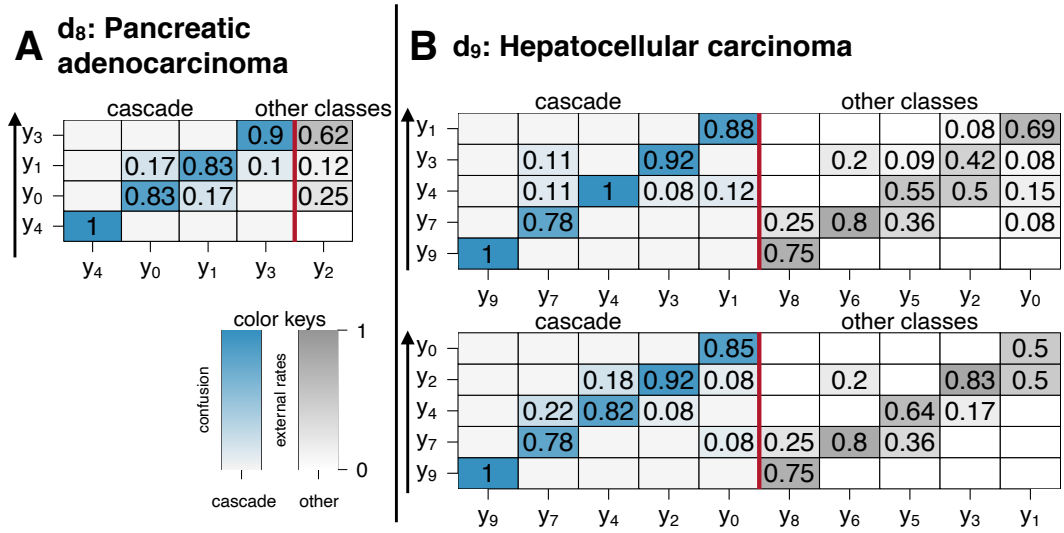


Figure 38: Classification of external classes for the returned pre-cancer cascades. For A: d_8: pancreatic adenocarcinoma (PDAC) Microarray dataset and B: d_9: hepatocellular carcinoma (HCC) dataset not all classes were included in the final selected subcascades. Their assignment to intra-cascade classes is given by the external rates. The classes representing the subcascades are given vertically from bottom to top. The left part of each heatmap represents the performance of the cascade with the class sensitivities on the diagonal. The right part shows the assignment of the external classes.

5.3.4.2 Gene set analysis

To gain more insight into the relation between the returned candidate set and the underlying cancer type, I analysed the feature sets for each cascade in the refined set that were selected by the linear L1-SVM embedded feature selection mechanism and using hOcc. There was a small number of features (5-17 %) selected robustly in more than 90 % of the folds of the leave-one-out-cross-validation (LOOCV) over all cascades (Figure 39). The overall number of genes depended on the cost parameter. For d_9 (HCC), with 10 classes, the, on the base classifier separability optimised, cost parameter of 6 resulted in an overall larger gene set. This dataset showed also the weakest performance on the level of pairwise separability (Figure 37).

The cost parameter of 0.9 for d_8 (PDAC, Microarray) resulted in 165 features with 28 selected in more than 90 %. In two analysed cascades there was a trend observed indicating more robustly selected genes with later positions of the cascade (Figure 39).

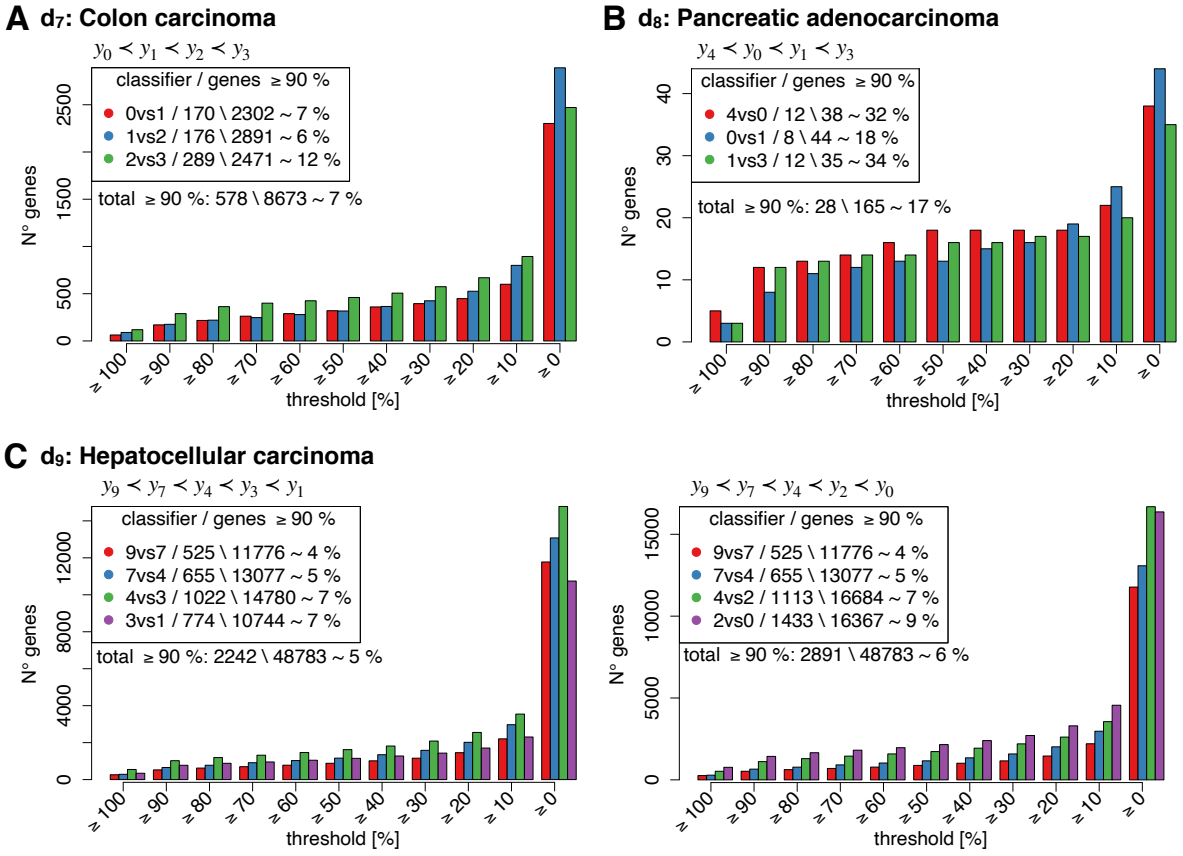


Figure 39: Histogram on the selected features for each finally selected cascade of the pre-cancer datasets. Features were considered as robustly selected if they were chosen in at least of 90 % of all folds. The total number of genes as well as the number and percentage of genes that are robustly selected is given at the bottom of the legend. The same measures are given for each base classifier in the legend. For the total number, features that are selected in more than one base classifier are counted only once.

Within all analysed cascades the selected features showed an overlap for all pairs of groups, except for the cascades returned for d_8 (PDAC Microarray) (Figure 40). In case of d_8 the features selected by $c_{y_1 y_3}$ were exclusively selected. For all remaining cascades of d_7 and d_9 there was an overlap of 3, 6 and 22 features selected by all binary classifiers. Nevertheless, the largest groups were always those features that belong exclusively to one step.

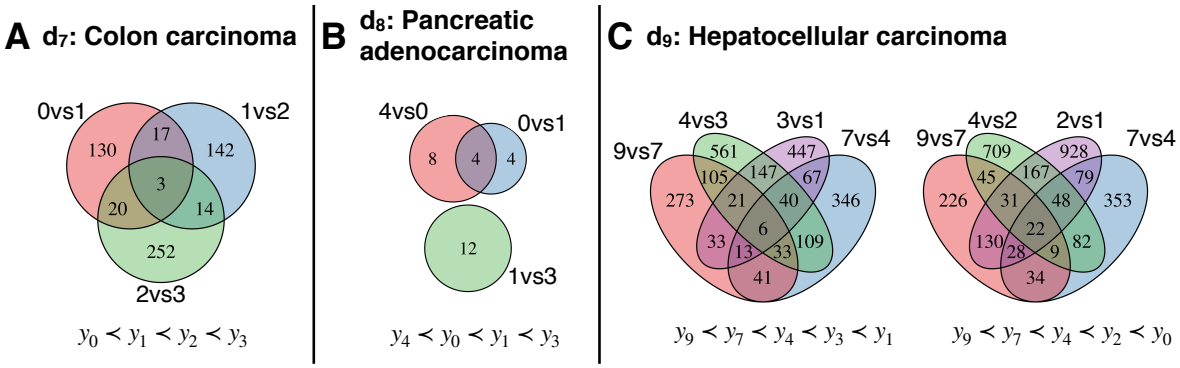


Figure 40: Overlap of the selected features of the base classifiers that make up the finally selected cascades of the pre-cancer datasets. The number of the robust features (selected in $\geqslant 90\,\%$ of the folds) per base classifier (colour coded) that make up the finally selected cascades for A: d_7: colon carcinoma, B: d_8: pancreatic adenocarcinoma (PDAC) Microarray, C: d_9: hepatocellular carcinoma (HCC) is shown.

5.3.4.3 Confirming carcinoma type relatedness

The relation between the selected features and the corresponding cancer type was analysed by performing a gene set enrichment analysis and literature search for d_7 and d_9. The small number of genes for d_8 PDAC was analysed via literature research only. It could be seen that all sets were enriched for genes belonging to pathways relevant for the corresponding cancer type. Among the top 15 pathways there were in case of the colon carcinoma dataset (d_7) pathways, that have not been found in gene enrichment analyses of the screened publications (insulin secretion, endocrine and other factor related calcium reabsorption, thyroid hormone synthesis) and pathways that have been detected in previous gene enrichment experiments and are well-known to be relevant for this cancer type, like wnt signalling pathway (Table 9). The same behaviour was observed for the HCC (d_9) (Appendix Table A3).

Table 9: Results on Kyoto Encyclopedia of Genes and Genomes (KEGG) gene enrichment analysis of the colon carcinoma dataset (d_7). The top 15 results from 27 results with an false discovery rate (FDR) adjusted p-value<0.001 are reported and compared to results from other studies. As characteristics the number of proteins within each pathway N°, the number of proteins included in the set (hits) and the FDR adjusted p-value are given. Additionally, the terms' id and descriptions and references to publications that have found the same pathway by performing a gene enrichment analysis on their sets of differential expressed genes are denoted.

term id	N° proteins	hits	adjusted p-value	term description	reference
d_7 Colon carcinoma $y_0 \prec y_1 \prec y_2 \prec y_3$					
1100	1161	37	7.7E-07	Metabolic pathways	[256, 172, 202, 213]
5200	320	17	9.6E-06	Pathways in cancer	[256, 255, 118, 202, 213]
4972	90	9	4.3E-05	Pancreatic secretion	[256, 250, 71, 139]
4110	121	10	4.3E-05	Cell cycle	[256, 131, 172, 71, 213]
4360	123	10	4.3E-05	Axon guidance	[213]
3010	128	10	5.2E-05	Ribosome	[250]
4911	81	8	8.6E-05	Insulin secretion	-
4970	82	8	8.6E-05	Salivary secretion	[125]
3013	151	10	1.5E-04	RNA transport	[172]
3040	124	9	1.8E-04	Spliceosome	[172]
3320	68	7	1.8E-04	PPAR signaling pathway	[255, 138, 172, 213, 139]
4961	46	6	1.9E-04	Endocrine and other factor-regulated calcium reabsorption	-
4918	70	7	1.9E-04	Thyroid hormone synthesis	-
4310	135	9	2.8E-04	Wnt signaling pathway	[256, 172, 213, 139, 125]
4114	105	8	2.8E-04	Oocyte meiosis	[213]

Many features selected in d_8 (PDAC, Microarray) relate to deoxyribonucleic acid (DNA) constructs that do not represent one gene. Those features that could be related to genes are given in Table 10. The genes that were selected in $c_{y_4 y_0}$ are well known to be related to PDAC and also some genes (progastricsin (PGC), phosphatidylinositol glycan anchor biosynthesis class Z (PIGZ)) selected in steps corresponding to pancreatic intraepithelial neoplasm (PanIN) progression have previously been reported (Table 10). Other genes, like death domain associated protein (DAXX), have only been reported in other cancer types.

Table 10: Robustly selected features in the pancreatic adenocarcinoma (PDAC) cascades (d_8) that represent genes. The ID is given once as feature identifier and once as gene symbol. Furthermore, it is denoted which feature is selected by which base classifier ($c_{y_4y_0}$, $c_{y_0y_1}$, $c_{y_1y_3}$), with 1 indicating "selected" and 0 indicating "not selected". Occurrences on the ribonucleic acid (RNA) and protein level are denoted. The statements in the database column are taken from a pancreatic cancer database (http://pancreaticcancerdatabase.org/). In addition to these statements further findings and their references are given.

ID		Base classifier			Database	Findings	References
identifier	symbol	$c_{y_4y_0}$	$c_{y_0y_1}$	$c_{y_1y_3}$			
NM_014405	CACNG4	0	0	1	-	-	-
AJ406940	partial KRTAP4.8	0	0	1	-	-	-
NM_033059	KRTAP4.11	0	0	1	-	-	-
NM_001350	DAXX	0	0	1	-	Overexpression reported in various cancer types, like prostate, ovarian, gastric, squamous cell carcinoma. It might have oncogenic properties.	[142]
NM_000788	DCK	1	1	0	RNA: slightly increased in Cancer [152]	Low level reported with age, which shows a positive correlation with survival. Expression is associated with Gemcitabine efficacy.	[94, 238]
NM_001929	DGUOK	1	0	0	RNA : increased in PDAC [36]	-	
AJ298844	DNASE1	1	0	0	RNA: increased in PDAC [36]	Dnase I treatment might decrease cancer metastasis	[229]
NM_002141	HOXA4	0	1	0	-	Has a potential oncogenic behaviour. It seems that it is not upregulated in cancer.	[117, 78]
AB058708	KIAA1805/ ZNF512	1	0	0	RNA: increased in PDAC [36]	-	-
NM_006121	KRT1	1	0	0	RNA: increased in PDAC [36]; Protein: increased in cancer [73] and pancreatitis [165]	Reported to be associated with survival time. It is underexpressed in tumours of patients with survival times >=10-year.	[44]
NM_002391	MDK	1	0	0	RNA: increased [211, 163]; Protein: increased [163]	Differentially expressed in histological subtypes (up). The upregulation is associated with worse overall survival.	[180, 252]
NM_002630	PGC	0	1	0	RNA: increased in PanIN1A, 1B [36, 175], in Panin2 [175], in invasive PDAC [103]	Differentiated tumour cells show a strong expression, whereas tumour stem cell show a low expression. It is positively regulated by c-MYC.	[30]
NM_003558	PIP5K1B	1	0	0	RNA: increased in PDAC [36]	Is upstream of KRAS and depletion can increase sensitivity to MAPK inhibitor in cell lines.	[5]
NM_005396	PNLIPRP2	1	1	0	RNA: decreased [79]; Protein: decreased [165]	High expression reported in the exocrine-like subtype. Found as decreased in tumour tissue compared to normal.	[60, 249]
NM_006917	RXRG/ NR2B2	0	0	1	-	Expression shown in pancreatic cancer cell lines and PDAC. RXR receptor hypothesised as acting as tumour suppressor but also reports of pro-oncogenic behaviour.	[174]
NM_003075	SMARCC2	0	1	0	-	Part of the SWI/SNF complex which was reported as tumour suppressor.	[82]
NM_003123	SPN/ CD43	1	0	0	RNA: increased in PDAC [36] and endocrine neoplasm [143]	CD43 was stained positively in PDAC and several other cancer types.	[23]

5.4 Discussion and conclusion

I screened through diverse real world datasets. These datasets can be divided into three groups: datasets with static classes, datasets comprising classes taken at different points in time during developmental processes and datasets comprising precancer stages with a hypothesised time relation.

Static datasets

The revealed relations within the datasets of static classes suggest that potential relations can indicate a variety of relations of various meanings. Within the HPA-axis dataset (d_1) one pattern in the set of returned cascades was hypothalamus $\prec$ pituitary $\prec$ adrenal. This is known as HPA-axis and corresponds to a signalling relation process known for stress reaction control [195]. As from liver to adrenal glands both directions were returned, the swap between adrenal glands and pituitary in one additional cascade is a swap in the last two classes, indicating a potential false positive result supported by the analysis of the artificial data (Section 3.3). Not considering this swap, the found set can be reduced to two potential sequences (liver $\prec$ skin $\prec$ hippocampus/hypothalamus $\prec$ pituitary $\prec$ adrenal). The similar methylation pattern in skin compared to the HPA-axis reported by Agba et al. [7] might explain the found neighbourhood. Furthermore, the interconnection between hippocampus and hypothalamus found by Rajmohan and Mohandas [177] and Jacobson and Sapolsky [100] might lead to the observed result that in stress response they cannot be placed into the HPA-axis by conserving their discriminability.
Analysing the leukaemia dataset (d_2) the interpretation of the relational axis is not clear. Nevertheless, observations can be made that might help to better understand similarities and differences of subgroups of cancer types and might guide knowledge transfer. First of all, in both subcascades a grouping can be observed into ML or syndromes that might develop into a myeloid leukaemia and LL. The two groups were separated by the class acute myeloid leukaemia (AML) with inv(16)/t(16,16). The exception is the non-leukaemia and healthy bone marrow class. For one of the subcascades this grouping could also be observed in the assignment of the external classes. This grouping showed a closer neighbourhood of distinct subgroups sharing a common context. One additional observation is that there is a high incidence for the expression of myeloid antigens in pro B-acute lymphocytic leukaemia (ALL) and early T-ALL [232], those subtypes found at the border of the ML-LL grouping. This leads to the hypothesis that the found relational axes might be interpreted as similarity axes that order well-distinguishable subtypes and the assignment of the external classes can give a hint for which subtypes knowledge

transfer might be possible.
Overall this screening has firstly shown the applicability of Occ towards screenings of large datasets with more than 15 classes and 50000 features. Secondly, it revealed relational axes that might be structural or similarity axes. These axes might improve the understanding of inter-class relations and serve as starting point for hypothesis generation.

For studying cancer development I was explicitly interested in relational axes possibly showing a time relation. In a broad sense changes in time might either happen predominantly due to well-controlled, deterministic or due to uncontrolled, stochastic processes.

Datasets of developmental processes

One example for predominantly well-controlled processes are developmental processes and many of the genes involved in these processes are evolutionary conserved [159]. By studying developmental processes of three model organisms (d_3 *Drosophila melanogaster*, d_4 *Danio rerio*, d_5 and d_6 *C. elegans*) I could show that the cascades included in the final candidate set corresponded to the assumed time relations. Consequently, I showed that the time relation is reflected in the gene expression data. The splitting into points in time instead of stages for the *C. elegans* dataset furthermore revealed that a too fine granular class definition leads to a loss in separability and hence sensitivity. Imposing a too high threshold subcascades instead of a total cascade might be found, which however still indicated an overall time relation, as the external classes were always assigned to their neighbouring classes.
The analysis showed the reflection of time relations representing developmental processes in the gene expression data.

Datasets of pre-cancer stages

Contrary to developmental processes cancer development is better characterised by stochastic disturbances of genetic information causing a high heterogeneity [81]. Furthermore, histological observations are based on tissue material taken when the patient underwent surgery and hence no time point sample taking experiments can be performed. As consequence potential pre-cancer stages are not defined based on their time point of occurrence, as this is unknown, but based on their histological appearance and the amount the tissue differs from normality.

Potential relations indicating an overarching relation between pre-cancer stages have to be revealed based solely on measurements and observations. This has been done based on histology. Pathologist have for example ordered the stages based on their grade of tissue disruption [114].
Whether this overarching relation is reflected in the gene expression data is assumed but so far there was no method to control this. Using the Occ I screened for potential orders in three types of cancer in order to search for a sequence of ordered and distinct stages. As there is an additional interest in the genes that might be relevant for a potential time order I used on the one hand the linear L2-SVM in concordance with the previous analysis and additionally the linear L1-SVM, which performs an intrinsic feature selection. As shown by Nilsson et al. [160] and confirmed by my cost parameter screening the linear L2-SVM does not depend on the chosen cost parameter. This also explains the robustness of the result on the type of base classifier parameter optimisation procedure.
Unlike, the results of linear L1-SVM as base classifier differed depending on the cost parameter and consequently the chosen type of base classifier optimisation. Nevertheless, robust solutions could be found and the general trend in the orders for the colon carcinoma (d_7) and the HCC (d_9) data was independent of the optimisation method, cost parameter, or whether one selects the best performing longest cascades or the Pareto optimal solution (Appendix Tables A1, A2). Neglecting a potential swap in the last two classes, exactly the same cascades of length 4 were returned for all colon carcinoma screenings and also the HCC resulted in the same set of best performing longest cascades for linear L2-SVM hOcc and linear L1-SVM hOcc. This was not the case for d_8 (PDAC, Microarray), resulting in two completely different orders for linear L2-SVM hOcc and L1-SVM hOcc.
The individual parameter tuning for the binary classifier might lead to an Occ with varying cost parameters and performance improvement on the binary level. However, only once classification performance improvement on the cascade level (best longest cascade d_7, aOcc) was observed, indicating that a homogenous cost parameter might be beneficial.

The longest cascades of the linear L1-SVM hOcc that are further analysed were found in at least 2 out of the 3 optimisation techniques for all three datasets, with the exception of $y_9 \prec y_7 \prec y_4 \prec y_3 \prec y_1$ for d_8. For two of the three datasets they were also returned by the linear L2-SVM. Screening through 37 cost parameters covering a range of 0.1 till 100, they were always included into the candidate set more often then 10 times, revealing their robustness and underlying the observed trend.

A detailed look at the set of returned cascades shows that for the colon carcinoma dataset (d_7) one total order was returned corresponding to exactly the proposed time relation. This however cannot be interpreted as for example adenomas always proceed to carcinomas. Many adenomas do not progress to cancer but many cancers develop from adenomas [34]. Instead, the found relation might also be interpreted as an order showing the grade of tissue disruption. As a result it can be said that my analysis revealed a reflection of the proposed order on all histological stages in the gene expression data for d_7.
Robustly selected genes ($\geqslant$ 90 %) at each step of the cascade of d_7 comprised neurexophilin and PC-esterase domain family member 4 (NXPE4) and X inactive specific transcript (XIST). Both genes have been reported in the context of colorectal cancer before and have been proposed as potential prognostic markers [242, 134]. XIST is a long non-coding RNA silencing one of the pair of X chromosomes. It has been shown to be involved in the chemoresistance to 5-fluorouracil (5FU)-based treatment of colorectal cancer [236] and might serve as a prognostic marker as its expression is associated with the overall survival rate [242]. The function of NXPE4 is unknown but it seems to play a role in chronic inflammation of the colon [241]. Furthermore, a significantly lower level in colorectal cancer, with an expression level that could be associated with tumour size and location, was found and hence NXPE4 was proposed as a prognostic maker [134]. The role of a prognostic marker is supported by this analysis, as these genes seem to be involved not only in progressing to a next stage but in forming an overarching pattern.
The overall gene set was enriched for genes involved in colon cancer related pathways, like Wnt signalling pathway or cell cycle. This combines the relevance of the selected genes with the interpretation of the found relation as showing cancer development. Insulin secretion seems not been found often in Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses but is known to play a role in colon carcinoma development [75, 246]. A similar aspect holds for thyroid hormone synthesis as hyperthyroidism correlates with cancer prevalence and thyroid hormones and receptors are assumed to promote proliferation in diverse cancer types, among them colorectal cancer [133]. The genes of the KEGG pathway endocrine and other factor-regulated calcium reabsorption have been found enriched in other illnesses and cancer types, like postmenopausal osteoporosis [240], renal cell cancer [153], pancreatic cancer [223] and prostate cancer [87].
Insulin secretion and thyroid hormone synthesis have not been found as detected by gene enrichment analysis of normal versus tumour tissue, but they are known to be involved in cancer development and progression. Hence, one might conclude that selecting genes under the constrain of an order between all stages can reveal pathways that are explicitly involved in the relational axes rather than in the aspect of best distinguishing the two extreme stages. Under this assumption it might be

interesting to look into the relation between colon carcinoma development and the genes of the endocrine and other factor-regulated calcium reabsorption KEGG pathway in future.

Similar observations can be made for d_9 (HCC). This dataset is divided into stages with a fine granularity and some cannot even be separated binary with a high sensitivity and specificity. As binary separation is a prerequisite for finding cascades it is not unexpected that no total order was returned comprising all 10 classes. The analysis of the external classes revealed that the returned orders do not contradict the proposed order of time relation, as external classes were assigned mostly close to their presumable neighboured stages. Solely samples of the carcinoma in situ stage were with a minority also assigned to the high inflammatory stages (high grade chronic hepatitis and cirrhosis) indicating a potential similarity between these stages. It might be hypothesised that at this stage inflammatory processes dominate.
Both selected cascades shared their top three ranked pathways within the gene enrichment analysis. Similar to the colon carcinoma dataset most pathways could be related to previous gene enrichment analysis of HCC by literature research (Table A3). The hypoxia-inducible factor (HIF)-1 signalling pathway is associated to the anaerobic metabolism and has not been found as enriched in any of the considered previous studies. Nevertheless, the role of hypoxia in HCC is widely studied. It is assumed that hypoxia activation protects tumours from radiation and high level HIFs in HCC are associated with poor prognostic outcome. Furthermore, a positive effect of HIF1α on hepatitis-C virus replication and the migration of hepatocytes could be shown and a potential pro-fibrotic function is hypothesised [42, 104], indicating that HIFs also play a role in the progression from fibrotic towards cancer stages. This is supported by the findings that genes involved in fibrosis are found enriched for that pathway [203].
One can conclude that the HCC (d_9) analysis supports the hypothesis that taking a progression order into account might reveal pathways and genes that are relevant for the actual progression via the pre-cancer stages. Opposed to a differential gene expression analysis, which concentrates on the differences between binary stages, ignoring the potential stage neighbourhood. Hereby, the selection contains most likely potential pathways that can be addressed to stop or delay the progression as they do not only differ between tumour and normal stages but are relevant for the overall relation between tumour and normal stages via the precursor lesions.

In contrast to the analysis of colon carcinoma and HCC the analysis of d_8 (PDAC, Microarray) did not show the proposed time relation neither for the linear L2-SVM Occ nor for the linear L1-SVM hOcc. It was robust towards these types of base

classifiers, as well. The normal healthy class was placed second with once the PDAC stage and once PanIN3 placed in front. The optimal homogenous cost parameter was small, resulting in a set of 165 selected features with 32 features selected in over 90 % of the cross-validation folds. This small gene set comprised many features (14/28) that cannot be mapped to genes, but also 14 genes that are known to be associated with PanIN stages or PDAC either on the RNA or even on the protein level. Due to the small gene set size no enrichment analysis was performed but the individual genes have been analysed.
The robustness towards the analysis parameters, the placement of the normal stage with stage PanIN3 and PDAC on opposite sites, the selection of PDAC related genes as well as the whole characterisation and validation results of the Occ and the previous results on real data indicate that the observed cascades were reflected within the data and were not a result of a random selection process. As consequence defining the root as healthy stage the feature space revealed a split structure with two opposite ways. On the one site there is the step from normal towards PDAC and on the other site the progression across the PanIN stages. This does not mean that PDAC cannot progress via the PanIN stages, it can rather be interpreted as under the assumption of one relation axes PDAC does not show the characteristics relevant for the progression axes via PanIN stages anymore. Most likely the observed results show the effect of chromotripsis, a massive rearrangement of the chromosome. Chromotripsis is more often found in PDAC compared to colon carcinoma or HCC. In PDAC around 75 % of tumours were polyploid or showed chromotriptic characteristics [162]. Imposing the definition of chromotripsis by Korbel and Campbell [113], others report 63 % of chromotriptic positive cases for PDAC and none for colon carcinoma [137]. Furthermore, they report 5.7 % and 25 % of chromotriptic positive cases, imposing less strict criteria of chromotripsis, for HCC in two independent studies [137].
The complete collapse and remodelling of the genome might lead to a different phenotype that do not share any progression characteristics with the remaining classes, explaining the location of y_4 and the y_3 on opposite sites of the normal class y_0. Genes selected for the progression steps are CACNG4, KRTAP4.8, KRTAP4.11, DAXX, DCK, HOXA4, PGC, PNLIPRP2, RXRG/ NR2B2, SMARCC2. Among them the first three have not been reported in context with PDAC. CACNG4 is the L-type calcium channel voltage dependent gamma subunit 4, involved in calcium channel regulation in excitable cells but also expressed in non-excitable cells. It is potentially metastasis related in breast cancer and might be targeted by calcium channel blockers [106]. Furthermore, this gene is involved in the mitogen-activated protein kinases (MAPK) signalling pathway and hypothesised to play a role in squamous cell carcinoma of urinary bladder (SCCB) progression [253].
Keratin-associated proteins could also be related to breast cancer, playing a po-

tential role in simplifying erb-b2 receptor tyrosine kinase 2 (ERBB2) amplification [147]. ERBB2 itself seems to be related to pancreatic cancer development [194, 47] and hence it is not unlikely that these genes also play a role in PDAC progression.

All in all this analysis supports the doubts in one clear progression for PDAC that was also mentioned by Real [182], Notta et al. [161, 162], Real and Andrés [183] and shows that there is no unique progression order from normal to the cancer stage in the analysed gene expression data.
The additional analysed RNA-Sequencing data revealed only cascades of length 3, and cascades of this length are found easier by chance. The very small sample size of this dataset and the observation that often several permutations of a subgroup of classes were found in addition to the fact that though the data analysis pipeline is almost the same for all classes, the PDAC data was generated by a different laboratory makes the results on this data possibly unreliable. To confirm the observed split in progression classes on the one hand and the actual PDAC class on the other the acquisition of more data is needed.

6

NETWORK ENTROPY HIGHLY DEPENDS ON THE PROTEIN-PROTEIN INTERACTION NETWORK

In the previous chapters I tackled the task of finding an order in pre-cancer stages by using the potential of the ordinal classifier cascade (Occ) to filter out not reflected orders in gene expression data. Another measure for distinguishing phenotypic stages and ordering pre-cancer stages might be network entropy [205]. The combination of gene expression with the information of the underlying gene network topology might have a beneficial effect on the distinction of stages. Additionally, it is hypothesised that it can directly be used to reveal an order between stages of varying differentiation status. It could be shown, for example, to systematically change while differentiation [18] and cancerisation in certain cancer types [46].
Two types of network entropy are distinguished: The local network entropy (LNE), considered as describing a non equilibrium state and the global network entropy rate (GNER), based on the stationary distribution and the largest strongly connected component (scc) of the network, for the equilibrium state.
However, network entropy is highly dependent on the network topology and protein-protein interactions (PPIs) measured by experimental and computational approaches are prone to produce false positives (FPs) and false negatives (FNs) [141, 29]. For example a false positive rate (FPR) of up to 64 % and a false negative rate (FNR) of 43 % to 71 % for yeast-to-hybrid screens, a common source of protein interaction knowledge, was claimed [116].

I analysed the influence of added, removed and flipped edges on the final entropy result of simulated microarray data, which is based on a yeast network often used as reference network [146]. Additionally, I tested several methods for correcting false positive edges. Finally, the selected strategy was applied to real data cancer sets $d_7 - d_{10}$.

6.1 Experimental setup

The network entropy relies on the network topology imposed. To test the influence of wrong connections I simulated 25 samples with the data simulator GeneNetWeaver [186, 146]. The simulator generates microarray resembling data using an automatic set up kinetic model based on differential equations and user defined PPIs as input. It is used to make in silico benchmarks [146]. Based on the DREAM4 challenge the integrated yeast PPI network (4441 proteins and 12873 interactions) was used as input PPI and the simulation settings were chosen. These settings are implemented as default in GeneNetWeaver [146]. The integrated yeast PPI is the true network underlying the simulated data and the data is considered as representing the wildtype phenotype.

The effects of gene expression changes and network perturbations on the true entropy, using the true network and the wildtype data as reference, were evaluated. Furthermore, the effect of various correction methods for FP interactions was analysed. A scheme of the experimental setup is given in Figure 41. Assessing the effect of the correction methods three types of deviations were distinguished. The systematic deviation describes the effect of the method on the true network and the stochastic deviation the effect the method shows on perturbations. The total deviation corresponds to the sum.

Effect of gene expression changes (knockouts)

In addition to the wildtype data, the simulator can handle knockout experiments by assigning a constant value of 0 as expression value to a modelled gene. Knockouts can be considered as gene expression changes potentially leading to different phenotypes. In order to estimate the scale of differences between phenotypes in this model, in silico knockout experiments on all single genes and on combinations of genes were performed. As simulating all dual knockout combinations would be computational unfeasible using the simulator, only genes within the largest scc were used for this experiment, as these genes are directly used to calculate the GNER and hence the strongest effect was expected.

Effect of network perturbations

The effect of network perturbations was evaluated by adding various degrees (1 %, 5 %, 10 %, 20 %, 30 %, 40 %, 50 %) of added, removed and flipped edges to this network. 100 independent repetitions per perturbation were performed and the impact of these perturbations on the true entropy was evaluated.

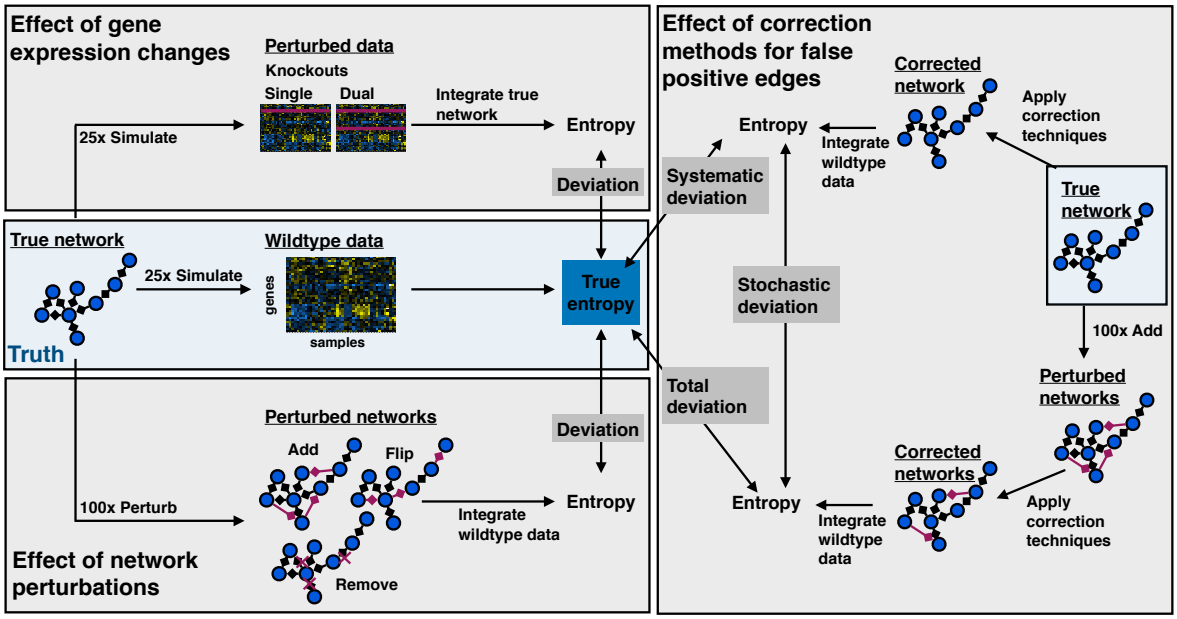


Figure 41: Scheme of the experimental setup for network entropy evaluation. On the basis of one true network, 25 samples were simulated that represent wildtype data. The entropy calculated based on the true network and the wildtype data was the true entropy used as reference (blue middle box). To estimate the effect of data perturbations the entropy was calculated using perturbed data and the true network (upper box). The effect of network perturbations was evaluated by adding, removing or flipping edges 100 times each and based on the simulated wildtype data (lower box). For added edges the effect of correction methods was measured (right box). These techniques were applied once to the true network leading to an estimation of the systematic deviation and once to the added perturbations. The corrective power of these techniques is measured by the use of the stochastic deviation.

Effect of correction methods for false positive edges

Using the true interaction network, the wildtype data and the added perturbed networks the potential of various PPI correction methods was evaluated.
In addition to the proposed methods, which filter edges according to a specific strategy, a random filtering out method was used as control. A number of edges were drawn uniform from all edges and deleted. The number of edges was determined by the parameter threshold, which corresponds in this case to the number of modified edges in percent. As too small networks might not be considered as representative, the network entropy was only calculated if the final network had more than 10 interactions and the calculation of GNER was furthermore filtered for more than 10 nodes.

6.2 Effect of gene expression and network perturbations on network entropy

The effect of network perturbations on the entropy was analysed with respect to the effect of gene expression changes. As network entropy might be used to distinguish phenotypes, the differences in entropy between phenotypes should be larger than deviations in entropy imposed by erroneous interactions in the PPI. To gain a reference phenotype difference, the scale of differences between phenotypes was analysed comparing wildtype data with knockout data. Knockouts are hereby considered as potentially representing another phenotype. The reference phenotype difference can be used to assess the scale of deviations resulting from network perturbations.

6.2.1 Effect of gene expression changes (knockouts)

Comparing dual and single knockouts the deviation from the wildtype was larger for dual than for single knockouts for both entropy types (Figure 42). Furthermore, it could be observed that the GNER showed about ten times larger deviations both for single and dual knockouts. Due to this larger deviation from wildtype the GNER might be considered as being more suitable for distinguishing phenotype stages.

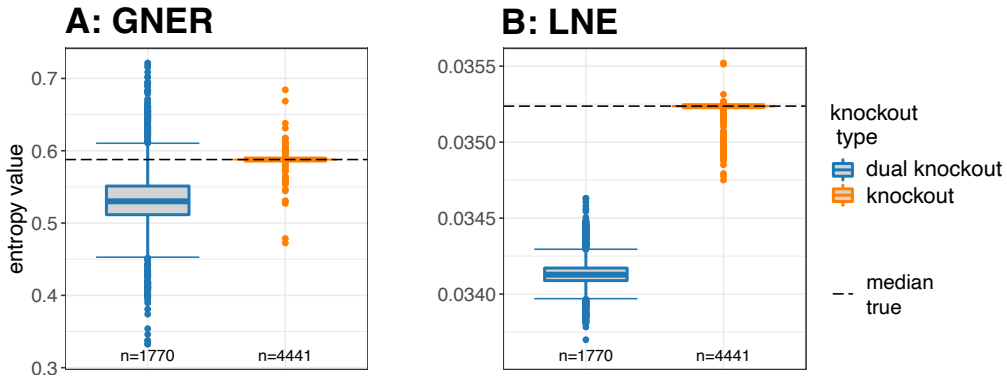


Figure 42: Network entropies of simulated phenotypes. The global network entropy rate (GNER) and local network entropy (LNE) of single and dual knockout experiments are denoted. The single knockout data shows entropy values of knockouts of each single gene and dual knockout experiments comprise all gene combinations of genes in the largest strongly connected component (scc). Each simulation was performed once without adding noise. The wildtype entropy values, as reference, are given as horizontal line.

In order to use a strict reference to judge the effect of perturbations on the network for each entropy type the maximal deviations resulting from a single knockout were used as reference phenotype difference. This resulted in deviations of approximately $(-0.196, 0.164)$ for the GNER and of (-0.014, 0.008) for LNE. Only if the deviation of a network perturbation was within these ranges the phenotype difference can be considered as being detectable, despite topological perturbations.

6.2.1.1 Effect of network perturbations

The analysis of the influence of the three different kinds of perturbations revealed a strong impact of flipped edges even at small noise levels (Figure 43). Removed edges seemed to impact the GNER the least, showing small deviations across all noise levels for the GNER and larger deviations only for the LNE and high noise levels. Added edges led to larger deviations on the LNE in comparison to the GNER. From a noise level of 20 % the deviation of the GNER would be too large to detect a single knockout in this setting and for the LNE this would be the case as soon as there is 5 % noise.

As flipped edges are not considered to be a common error type in a PPI and removed edges show less impact on the calculated entropy, I concentrated on added, hence false positive edges in the following.

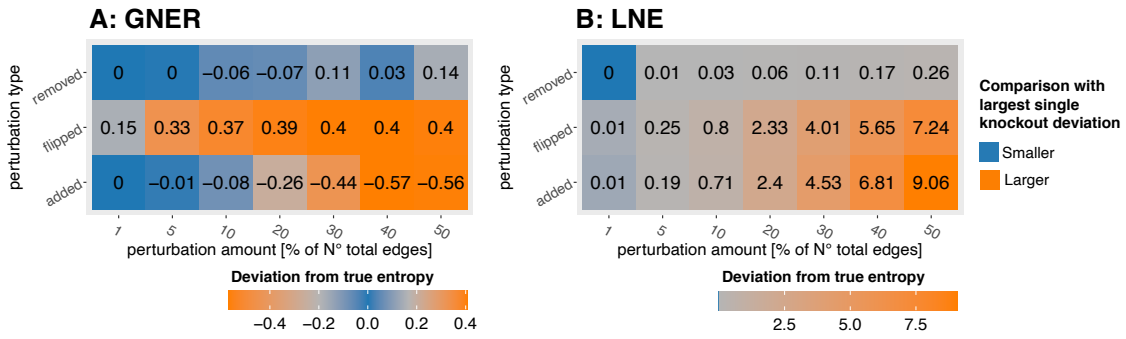


Figure 43: Impact of three perturbation types on the network entropies. The deviation of the true entropy for perturbation experiments with added, removed and flipped edges is shown for A: the global network entropy rate (GNER) and B: the local network entropy (LNE). Each perturbation was performed 100 times and the median deviation is depicted. The colour coding reflects the thresholds revealed by the single knockout experiments. A blue colour corresponds to a deviation smaller than the largest single knockout deviation. Orange colour indicates deviations larger than the reference deviation.

6.2.2 Effect of correction methods for false positive edges

Each analysed correction method might already show an effect on the unperturbed network by filtering true positive edges (systematic deviation). I have first analysed this effect before I measured their potential to correct FP edges.

6.2.2.1 Influence of correction methods on true network

In a first step the effect of each method on the true network entropy was evaluated without any permutations imposed. Figure 44 shows the deviation from the truth of each method and thresholding for four characteristics: GNER, LNE, number of nodes and number of interactions of the largest scc. It can be seen that the composite scoring by Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) led to a strong drop in the size (number of nodes and interactions) of the largest scc for all thresholds. Contrarily, this decrease could only be observed for larger thresholds in the remaining methods. Additionally, the effect of concentrating on networks exceeding a proposed minimal size could be seen, as for example already for small thresholds using Jaccard or Czekanowski-Dice (CD) scoring the corrected network became too small. Whereas, the LNE showed an overall decrease, this was not the case for the GNER, showing deviations in both directions. The composite scoring by STRING led to an overall overestimation for GNER.
Each method showed an individual optimal threshold in order to reach a minimal deviation for the analysed data and the systematic deviation should in general be kept in mind if a correction method on the PPI is applied.

6.2.2.2 Potential of correction methods for false positive edges

As false positive edges showed a high impact on the deviation from the true entropy, several methods were tested on their potential to correct for those. The deviation imposed by perturbations is measured by the statistical deviation. Figure 45 shows this deviation on the GNER for all methods and thresholds. The corresponding plot for the LNE can be found in Appendix Figure A9. The focus lied on correcting the deviation for noise levels larger 10 % as below the deviation resulting from network perturbations was smaller than the largest single knockout deviation, used as reference for phenotype differences (Figure 43). The methods are evaluated with respect to this phenotype reference deviation. For most thresholds all semantic methods did not perform well for these noise levels, as the resulting deviations were larger than the phenotype reference deviation, with the exception of Resnik and a threshold of 0.1. The topological methods performed similar for soft thresholding and a threshold of 0 but with increasing thresholds Jaccard and CD filtered out so many edges that they did not match the minimal requirement on network size anymore. The inverse log-weighted (invlog) method showed a performance

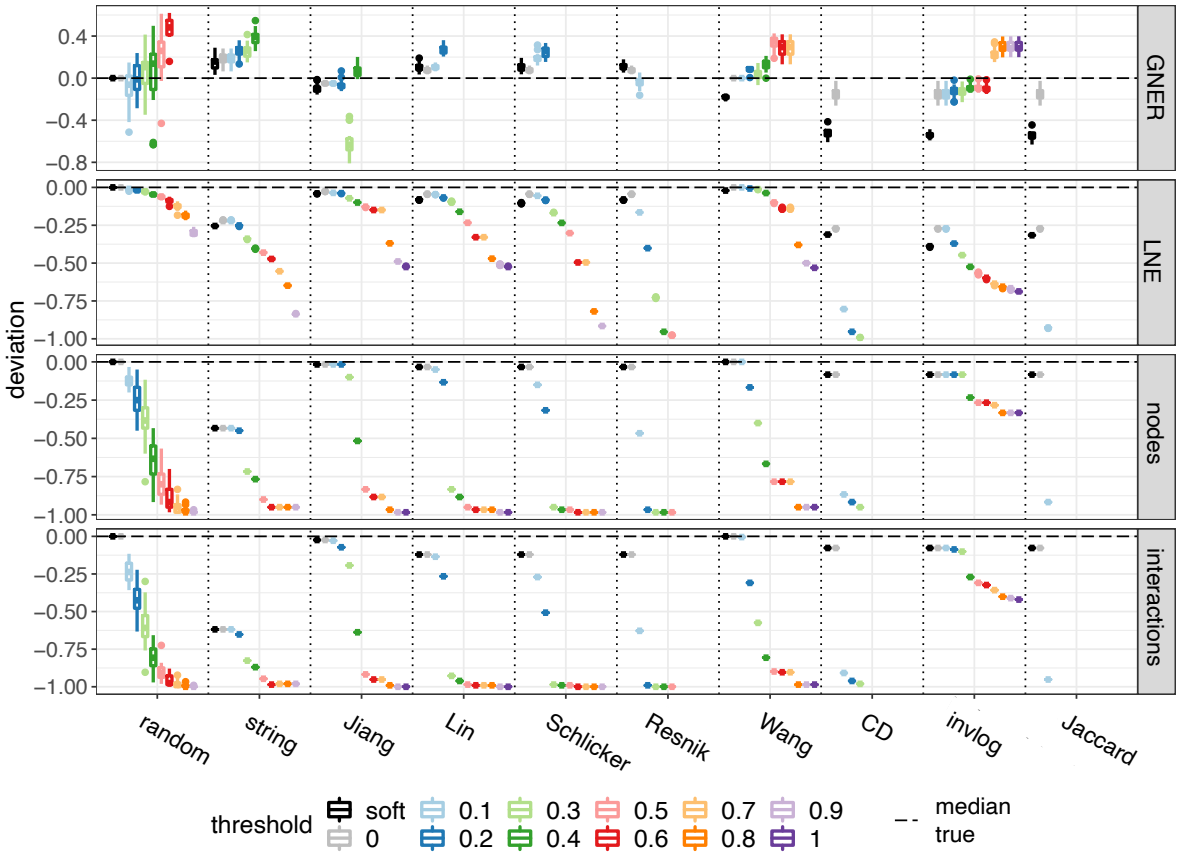


Figure 44: Correction technique influence on true network characteristics. The deviation from the true values based on the true network structure for four characteristics is shown in the rows: global network entropy rate (GNER), local network entropy (LNE), number of nodes and number of interactions of the largest strongly connected component. Each column corresponds to one correction technique with the individual thresholds colour coded. There are 25 samples simulated. Czekanowski-Dice (CD), inverse log-weighted (invlog).

optimum in the threshold range of 0.1-0.4 becoming worse for larger thresholds. Nevertheless, the string scoring outperformed the other methods, showing small deviations for all noise levels and all thresholds. A comparison with the random control revealed that the improvement was not due to a decrease in the overall network size but due to the systematic sorting out or weighting of edges. As also for the string scoring the network did not exceed the threshold on network size once the threshold was set too large, I propose to use a soft thresholding, fusing the information given by the reliability score with the expression values and by this handle the uncertainty of the edges rather than filtering those.

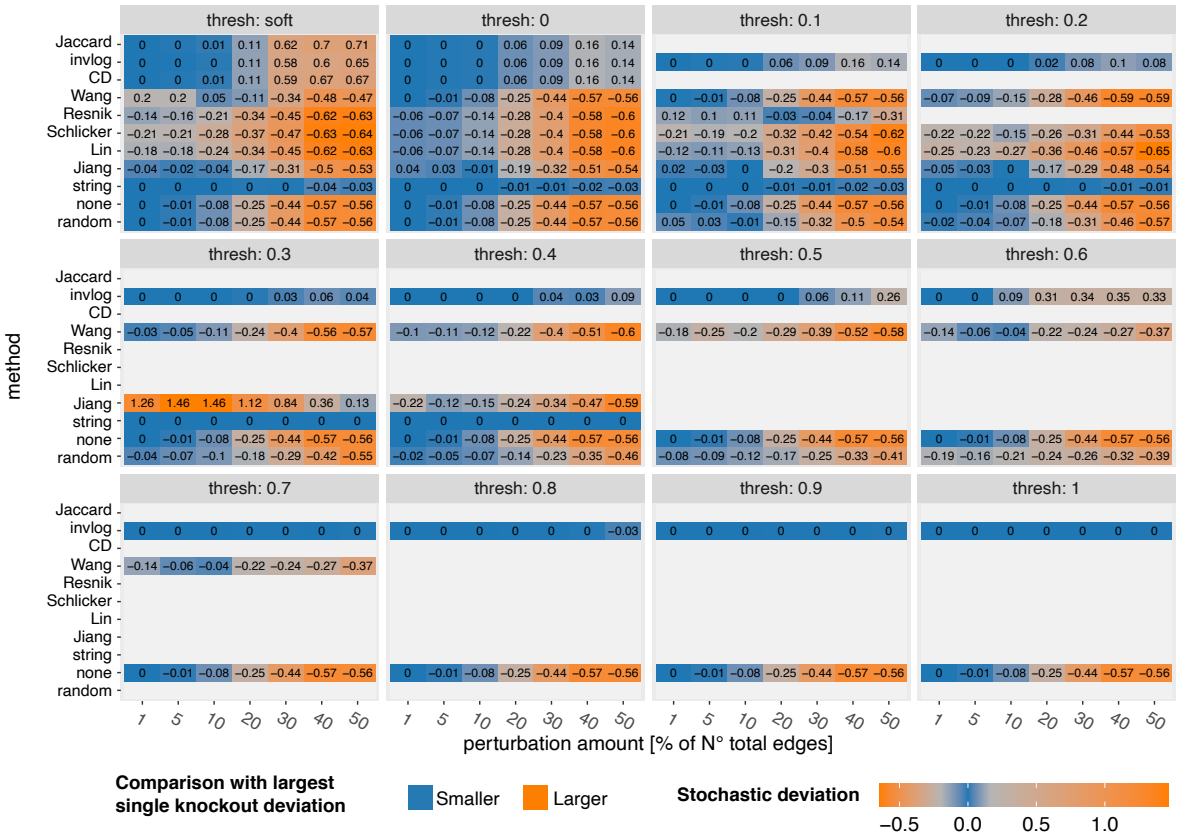


Figure 45: Performance of false positive correction techniques. The relative deviation between the global network entropy rate of the unperturbed corrected network and the perturbed corrected network is shown (stochastic deviation). Each panel corresponds to one threshold and comprises the various methods in the rows and the amount of perturbation in the columns. The colour coding is with respect to the maximal deviations measured between a single knockout and the wildtype stage. Blue colour corresponds to deviations smaller than the deviation between knockout and wildtype and orange to deviations larger than these thresholds. Czekanowski-Dice (CD), inverse log-weighted (invlog).

6.3 Network entropy courses of the (pre-)cancer stages

The proposed correction strategy was applied to the cancer datasets $d_7 - d_{10}$. The size of the largest scc stayed constant across all classes. For the colon carcinoma (d_7) and the hepatocellular carcinoma (HCC) (d_9) a trend towards higher GNER with increasing stages could be observed. However, the metastatic stage in the colon carcinoma showed again a smaller GNER and the HCC showed a small peak at the cirrhotic stage (y_4). This trend could not be observed in neither of the pancreatic adenocarcinoma (PDAC) datasets (d_8, d_{10}). Instead the microarray data revealed a decrease in entropy for pancreatic intraepithelial neoplasm (PanIN)2 and a

slight peak at PanIN3 with a decrease again for PDAC. The ribonucleic acid (RNA)-Sequencing data seemed to be similar but due to the high variance the stages might generally be considered as not distinguishable (Figure 46).

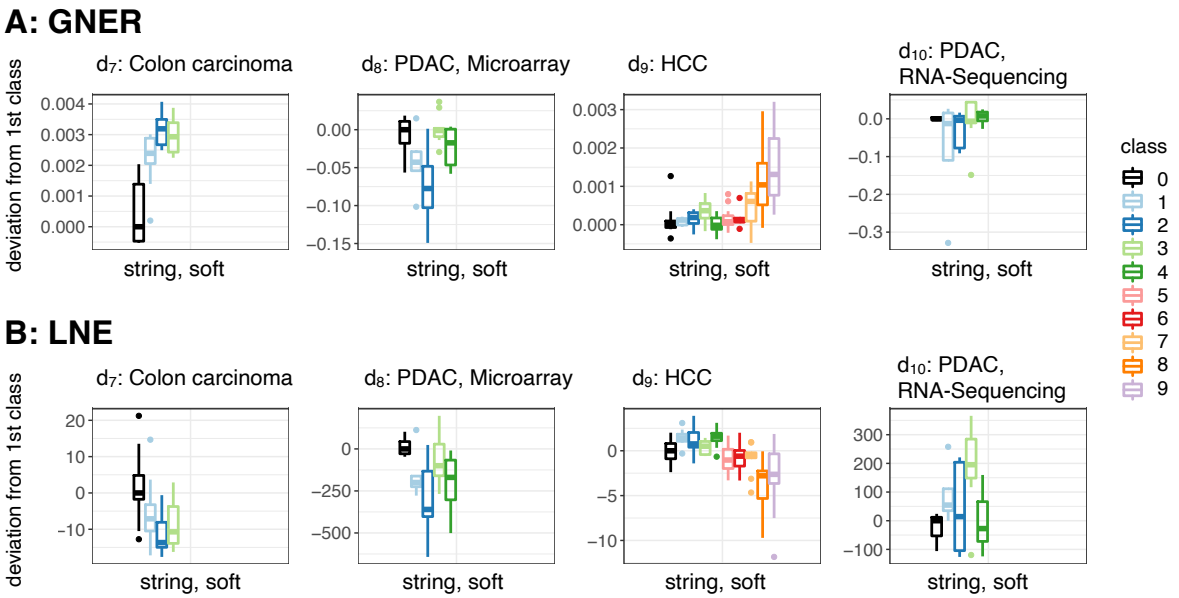


Figure 46: Entropy results on four pre-cancer datasets. The deviation from the first class for the global network entropy rate (GNER) and the local network entropy (LNE) is shown for colon cancer (d_7), for pancreatic adenocarcinoma (PDAC) measured once using a microarray (d_8) and once by ribonucleic acid (RNA)-Sequencing (d_{10}) and for hepatocellular carcinoma (HCC) (d_9). The median of the first class is used as reference for all deviations. Each dataset comprises a different amount of classes and each class is colour coded. The classes are ordered based on the hypothesised order on the stages and the network used is the union network of three databases (Biological General Repository for Interaction Datasets (BioGRID), Pathway Commons (PC), Search Tool for the Retrieval of Interacting Genes/Proteins (STRING)).

As I neither have paired nor normal distributed samples I tested the class differences for significance using a Mann-Whitney-U test statistic and applied Holm correction [89]. The significantly ($p<0.05$) differing pairs are given in Figure 47 and the corresponding p-values are given in Appendix Tables A5. It could be observed that for colon carcinoma (d_7) only the cancer compared to the metastatic stage was not significantly different using the GNER. Analysing PDAC the RNA-Sequencing data (d_{10}) revealed significantly differences for PanIN3 and PDAC using LNE. Within the PDAC microarray dataset (d_8) only the normal and PDAC stage were distinguished from each other using the LNE. For both datasets no significantly different pairs were found using GNER. In contrast to this there was a significant difference between TG2, TG3 cancer stages and some potential pre-cancer stages for HCC (d_9)

and the GNER. Interestingly, the LNE showed additional significant differences, also between pre-cancer stages.

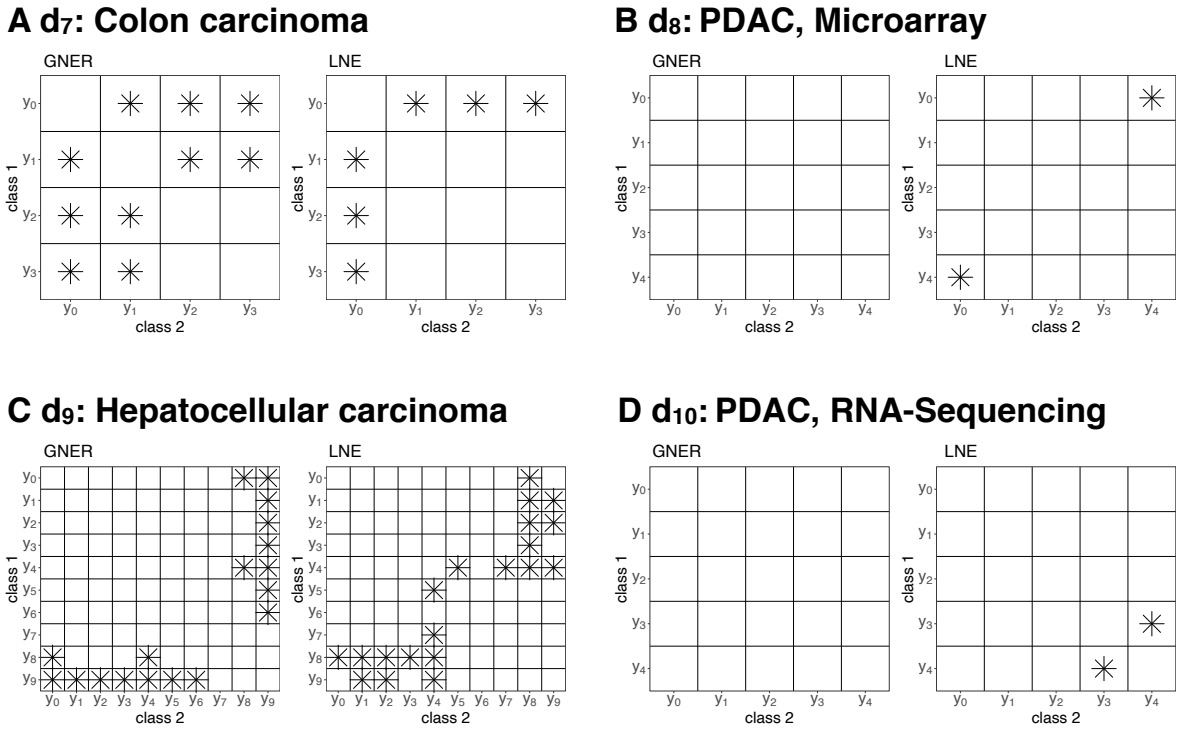


Figure 47: Significant pairs of classes using a Mann-Whitney-U test with Holm correction [89]. The class pairs that showed a significant difference ($p<0.05$) using a Mann-Whitney-U test are marked with an asterisk. The results for d_7: colon carcinoma, d_8: pancreatic adenocarcinoma (PDAC), Microarray, d_9: hepatocellular carcinoma (HCC) and d_{10}: PDAC, ribonucleic acid (RNA)-Sequencing and global network entropy rate (GNER) as well as local network entropy (LNE) are given.

6.4 Discussion and conclusion

Differences and relations between phenotypes might be investigated purely based on the measured gene expression levels. As the information flow between proteins is constrained by their abilities to interact further knowledge might be gained by incorporating static knowledge about the PPI into the analysis. One fused measure incorporating gene expression and protein interaction knowledge, is network or signalling entropy. Two types of network entropy are used to investigate phenotype differences LNE and GNER, with LNE being considered as describing a non equilibrium state [18]. Both measures are published as being suitable for distinguishing normal and tumour stages [230, 46, 18] and might be used as potential measure on the Waddington landscape. The findings are gained using self-built PPI networks that were constructed based on PPIs having been measured using diverse techniques and are collected and stored in a set of databases. However, PPIs stored in databases evolve over time, as some new interactions might be found and other interactions might not be confirmed using new techniques. But not only the network in one database differs depending on the evaluation time point also the databases differ among each other and additionally it is estimated that they contain many FP and FN interactions [116]. Furthermore, network entropy phenotype differences were shown to depend on the property of the PPI to be scale-free [204]. But so far the impact of topological perturbations on network entropy has not been measured.

Effect of gene expression and network perturbations on network entropy

The evaluation of entropy differences between phenotypes (normal versus single knockout and normal versus dual knockout) showed a higher impact on GNER than on LNE for the analysed network. This might be explained by the use of the largest scc for calculating the GNER. The simulated yeast network had many nodes that are set as input or output node and hence many nodes that showed an entropy of zero. The entropy of these nodes is not considered using the GNER but directly contributes to the LNE calculated as mean of all LNEs. Although this network is often used as reference network for evaluating PPI network reconstruction techniques, one might be careful to transfer the strong differences between GNER and LNE in the phenotype differences analysis to larger networks, which show a less strong difference between all nodes and nodes included in the largest scc.

The range of phenotype entropy differences allows to analyse and interpret the impact of entropy differences introduced by network perturbations. The impact of per-

turbations was always analysed with respect to the impact of the stricter criterium the single knockout, which showed a smaller deviation from wildtype compared to the dual knockouts. From the three perturbation types (added, removed, flipped) flipped edges showed the highest impact on entropy but are also considered as least realistic and are not found as common error type in literature. The strong impact might be explained by inversions of network flow. The impact of removed edges was for all noise levels and both entropies lower than the impact of added edges. For GNER the relative deviation was always lower than the corresponding difference between wildtype and knockout entropy. The lower impact might in case of the GNER be explained by many of the edges removed not being in the largest scc and hence having only an indirect impact. Nevertheless, the total percentage of removed edges should correspond to the removed edges within the largest scc. In order to correct removed (FN) edges one would need to predict edges.
For perturbations by adding edges the impact of the FP became larger than the impact of a knockout if 20 % or more FP edges have been introduced. This shows that under the assumption of an erroneous PPI and the inclusion of this variance some differences measured as significantly different might actually not be distinguishable.

Potential of correction methods for false positive edges

The influence of FP edges might be reduced by PPI correction techniques. All techniques introduced a systematic deviation on the network entropy and hence one might be careful in analysing and comparing absolute values. This systematic deviation can be explained by the elimination of actual true edges. In combination with the findings that the entropy deviation for small noise levels in added edges is relative low it might be better not to use any correction technique in these cases. Nevertheless, if one deals with PPI networks that are considered to represent the whole PPI network high noise levels are assumed and the question of the best correction technique arises.

Comparing the three strategies (topological, semantic, composite by STRING), the composite by STRING correction strategy is among the best techniques independent of the chosen threshold. The results for the different kinds of topological and semantic techniques are similar among each group, indicating that the general additional knowledge has a higher impact than the detailed processing of this knowledge. Furthermore, the comparison with the random control revealed that the corrective power is due to the imposed strategy and not only due to the elimination of edges.

For the thresholding techniques soft thresholding and hard thresholding was distinguished. The choice of the threshold strongly influenced the result. If it was chosen too high, the network broke down and was not considered as representative for the whole network anymore. If it was chosen too low no correction was performed. An optimal threshold might be found by using optimisation techniques. However, these techniques require an optimisation criterion, which cannot unambiguously defined in case of an unknown underlying true network. In case of soft thresholding one does not need to decide for a threshold and no general network reduction is performed. These aspects in combination with the good performance of the string strategy determined the choice of correction strategy and draws attention to the fact that all differences measured have always been analysed with respect to the used underlying network.

Network entropy courses of (pre-)cancer stages

Using the proposed strategy on the cancer data, the previously published differences between normal and tumour samples for colon and liver (d_7, d_9) could be reproduced [18, 204]. The findings by Sato et al. [185], who showed a systematic increase in entropy for cervical carcinoma via the pre-cancer stages, the findings by Banerji et al. [18], who measured a systematic decrease in entropy while differentiation and the hypothesis that cancer cells resemble less differentiated cells more and have a less distinct information flow leads to the expectation of an increase in entropy via the pre-cancer stages. This expectation was partly confirmed in case of GNER for colon cancer (d_7) and HCC (d_9). In case of colon cancer (d_7) the GNER only dropped for the metastatic stage. However, the same behaviour was observed by Cheng et al. [46], who analysed LNEs and it might be interpreted as the metastatic stage having a less distributed information flow. For HCC (d_9) the course showed an intermedian peak for the cirrhotic stage. One might argue that this stage is the end stage of a course of inflammation stages and that although HCC is proposed to develop via a a cirrhotic stage [148] it does not represent a pre-cancerous stage. As a trend could be observed but many stages were pairwise not significantly different this analysis might give a hint but the hypothesis that the actual start point of HCC progression is the dysplastic nodules stage needs further confirmation.

Analysing pancreatic cancer some similarities might be observed between the microarray data and the RNA-Sequencing data (d_8, d_{10}), like smaller entropies for PanIN2 and PDAC than PanIN3 for LNE and a potential downward curse towards PanIN2 for GNER. As, however, the sample size for the RNA-Sequencing data is very low and no differences for the GNER and only one for the LNE

were significant, it is almost impossible to interpret the results in a valid way. No significant differences between classes were measured in the PDAC microarray dataset and the expected course in GNER was not seen. Instead of an increased LNE and GNER with progression stages the entropy decreased until PanIN2, increased for PanIN3 and slightly decreased for PDAC. The slight peak for PanIN3 might go along with the findings by Hosoda et al. [90], who did not find mutations in SMAD4 in PanIN3 samples received from non-cancer patients and proposed that the PanIN3 stage comes after PDAC. In general, this analysis showed that PDAC progression behaves differently to the two other cancer types and further analysis on the decrease in entropy for PanIN2 and the relation between PanIN3 and PDAC might lead to a better understanding of the importance and relevance of the individual precursor lesions.

7

CONCLUSION

Multi-class datasets often comprise samples of various classes united under a common research context. Based on this limited amount of data one searches for concepts that employ specific data characteristics to define general rules that can then be applied to new data. One concept of interest is the relation between classes. Hereby, the classes correspond to a grouping and labelling of samples, and the samples are given by the measurements. These relations might be of interest in basic research to analyse development or relationships between organisms or phenotypes. But they are also relevant in medical diagnostics as they might guide treatment options or serve as basis for knowledge transfer.
If solely one feature is measured, an ordinal relation might be concluded directly. For example, in histological examinations of pre-cancer stage tissues only the grade of observable tissue disruption with respect to the normal stage is measured, and a semantic order on pre-cancer stages is proposed based on this measure. If in the following a molecular profile is gathered additionally, an order between classes cannot easily be defined due to the quantity of different features but the previously proposed order might be assumed to be reflected in the molecular profile. This assumption, however, has not been proven and requires further analysis.

Transferring semantic ordinal relations between measurements or searching for them in high-dimensional data goes along with the following questions:

- How to check whether a proposed linear order is reflected in the given measurements, how to reveal linear orders from them and how to combine several linear ordinal relations to one order structure?
- Which data structures should be considered as ordinal in high-dimensional space?
- How can the revealed ordinal relations be interpreted?

The third question, the interpretation of orders, is research context specific and in most cases will probably be answered with the aid of external domain knowledge or lead to hypotheses that are then further investigated.

The first and second question are closely related as the technique of analysis will determine the kind of order and vice versa. One possible idea of orders in high dimensions is exploited by the data depth concept by Tukey [212]. As the data depth describes the centrality of a sample with respect to the other samples, the concept leads to a center outward ordering of the samples. In this approach the class labels are not considered and there is neither a requirement on the nearby grouping of samples of the same class nor is a clear separability of distinct classes given. In contrast to this, the working definition of Bellmann and Schwenker [25] is based on one specific order on the class-wise accuracy matrix, resulting from pairwise classification tasks. As classification techniques are employed an order between clearly distinct classes is expected. However, based on Cover's theorem [51] this approach most likely suffers from the curse of dimensionality and is not applicable to high-dimensional data.

In this thesis, I target the first and second question by the functionality of the ordinal classifier cascade (Occ). In contrast to approaches that use ordinal classification solely to improve classification performance [25, 181, 173, 69, 68, 38, 65], ordinal classification is hereby rather used for data analysis purposes. The third question is targeted by the application of Occ to various real data examples, among them pre-cancer stages of different cancer types.

Introduced by Lattke* et al. [119] Occ, performance was shown to be sensitive to wrong order assumptions. As, however, the drop in performance for non-reflected orders only becomes visible after having screened through all order permutations, Occ is not applicable to large datasets with many features in a straight forward way. Within the scope of this thesis, together with colleagues, I introduced two improvements that now allow for the usage of Occ in exhaustive screens in large datasets (Section 3.1). Firstly, the employment of Theorem 1 allowed for thinning out the search space based on the performance of the pairwise trainings and reducing the amount of permutations that have finally to be evaluated. Secondly, saving the results of the pairwise trainings in a look-up table structure, a prediction map, further prevents a de-novo calculation of the performance measures in the final evaluation. Furthermore, I made the screening functions publicly available by bundling them together with a set of analysis functions into an R-package *ORION* (Section 3.1.4).
Using artificial data, I first characterised Occ with respect to its behaviour with increasing dimensionality and afterwards with respect to the orders that it returns. Interestingly, the results showed that the Occ does not suffer from the curse of dimensionality due to its exploitation of external rates during evaluation (Section 3.2).

The analysis of the orders returned in screenings of artificial data showed that Occ generally returns the expected orders. The artificial data was sampled in order to resemble orders on biological phenotypes and various relational patterns were imposed. However, I also detected two possible error types: First, a swap in the last two classes of a cascade; Second, cascades of length three might be more likely to be found by chance. The first error type might be explained by the least constraints at the end of a cascade, whereas the second might be a result of too few exploitations of external rates, as the shorter the cascade the less often samples of foreign classes have to be traversed within the Occ (Section 3.3).
In cases in which not one total order but several orders of subset of classes are revealed, the question of how to combine them to one order structure arises. This task is related to finding a partial ordinal structure in a given data representation. I proposed a strategy to combine (sub)cascades and showed that this strategy performed better than pure distance measures or self-organising maps for the assumed task in finding ordinal relations between phenotypes (Section 3.4).

The analysis of the artificial data already gave a first insight about the data structure that is found by Occ as ordinal using linear support vector machines (SVMs) as base classifier. However, in order to compare different kinds of orders, a theoretical background for types of orders in high dimensions was missing. Based on the definition of a component graph, built up by decision regions, I provided different categories, each describing a different aspect of what might be considered as ordinal. These categories might also be used to other ordinality approaches. The working definition of ordinality by Bellmann and Schwenker [25], for example, can be considered as fulfilling at least category III, as there are no disconnected decision regions. Furthermore, I showed that the impact of the base classifier characteristics (k-nearest neighbour (kNN) versus SVM) of Occ can be used to distinguish between the categories that allow disconnected decision regions and those that require connected decision regions (Section 4).

After having improved and characterised Occ as well as analysed different kinds of ordinality in high dimensional space, I applied the approach to three categories of real data. Within the analysis of static sets (Section 5.1), no order between all classes was hypothesised a priori. I interpreted the resulting orders by using domain knowledge. Based on the results I could not only show that the structural hypothalamic-pituitary-adrenal (HPA) axis is reflected in the methylation data but also hypothesise that the skin and liver might be potential additional elements. Furthermore, an overall neighbourhood structure on several subgroups of leukaemia was revealed. One subgroup seems to serve as connecting link between lymphocytic leukaemia (LL) and myeloid leukaemia (ML) in this structure. The

relations revealed might not only give a deeper insight into the research context, they can possibly be exploited for knowledge transfer as well.
In the analysis of developmental stages of model organisms, the expected order reflecting timing was found in all datasets (Section 5.2).
In contrast to these developmental processes, cancer is believed to be determined by stochastic perturbations rather than by controlled deterministic processes in time. Nevertheless, I found an order consistent with the proposed stepwise hypothesis for colon carcinoma and hepatocellular carcinoma. Furthermore, I figured out that discriminability of classes and length of subcascades depended on the granularity of the defined stages. Yet, even if only subcascades were identified, external classes were assigned to their presumable neighbouring classes and conclusions on an order corresponding to the chronological order could be drawn. To better understand and relate the found order to specific genes, I applied a linear L1-SVM in addition to linear L2-SVM to the same data. The gene enrichment analysis did not only allow to relate the found orders to the underlying cancer but also revealed pathways that might contribute to the progression of the tumour. Such pathways have not been described in the context of gene enrichment analysis before. This aspect could point towards an advantage of Occ to find an overarching pattern between several classes rather than pairwise differences alone. Conversely, no consistent order was found for pancreatic adenocarcinoma (PDAC).
The analysis of network entropy, a measure that might serve as ordinality measure in case of phenotypes [18, 46], supports the doubts on the ordering of pre-cancer stages with respect to PDAC (Section 6). Results gained from network entropy, however, have to be analysed with respect to the used protein-protein interaction (PPI). Extensive characterisation of network entropy supported that it highly depends on the PPI.

All in all, I could show that the task of confirming and finding ordinal relations within high-dimensional data can be tackled by the use of Occ, with the additional benefit that it does not suffer from the curse of dimensionality. Furthermore, the characterisation experiments showed that for reconstructing phenotype relations the proposed reconstruction strategy worked best within the set of analysed methods. The proposed categories of kinds of orders in high dimensional space might serve as categorisation framework for comparing and analysing different methods that reveal these orders. The analysis of real data finally showed that Occs can be used to find previously unknown ordinal structures, confirm hypothesised orders or else to support doubts of them.
The characterisation of network entropy, another measure that was proposed to be able to capture ordinality on phenotypes, revealed that this measure is highly sensitive to the used external knowledge (the PPI). All results therefore need to be analysed with respect and under consideration of the used PPI.

8

SUMMARY

When investigating more than two defined classes in a common research context, orders and relational patterns between these classes are often assumed to be reflected in the collected measurements that determine the feature spaces. For example, orders of cellular phenotypes might be defined based on histological observations and the same orders are also assumed to be reflected in their respective gene expression patterns. So far, however, this assumption has not been verified.
Furthermore, key aspects in medical diagnosis include a precise categorisation of pre-cancer stages as well as the knowledge of their nearby neighbourhoods. Both aspects serve as guidance for potential treatment procedures. With the possibility to gather high-throughput data as a basis for medical decisions regarding tumour treatments, there is need for a procedure that allows for the screening of orders between well-distinguishable classes in a given feature space.

In this thesis, I showed that by applying an ordinal classifier cascade (Occ) I was able to tackle both aforementioned tasks. Ordinal classification is hereby not used in its original sense, classification performance improvement [95], but for data analysis. Having improved the screening procedure, it can now be used for exhaustive screenings in collections of large datasets with many features. Additionally, bundling the screening and analysis functions into an R-package *ORION* makes it versatile (Section 3.1). Furthermore, I showed that Occ does not suffer from the curse of dimensionality (Section 3.2). Extensive research on artificial datasets with relational patterns imposed on the feature space showed not only that using Occ enables the finding of the underlying relations but also revealed two aspects that might be considered as weak points and should be taken into consideration (Section 3.3): A swap in the last two classes of a cascade, and cascades of length three are likely to be found by chance and prone to represent false positives. The swap might even occur using well-distinguishable categories and can be explained by the least constraints at the end of a cascade. Using artificial datasets with partial orders, I proposed a procedure for combining the returned (sub)cascades to reconstruct the partial order and could show that this procedure performs better

than pure distance measures or self-organising maps (SOMs), with the additional benefit that one deals with well-distinguishable categories (Section 3.4).

In general, the meaning and interpretation of ordinality in high-dimensional spaces is not straight forward. One might refer to Tukey's data depth concept to describe ordinality on a sample level [212] or to a constant increase or decrease of all features, as in monotone classification [65]. Nevertheless, on the sample and on the class level, many structures might be considered as ordinal and the interpretation requires domain knowledge. The theoretical background for analysing ordinality in high-dimensional spaces proposed in this thesis provides a basis for distinguishing, subgrouping and ranking types of ordinality. However, I developed not only the mathematical background, definitions and categories of ordinality. Furthermore, I showed that Occ can be used to distinguish between certain classes by using the impact of the base classifier characteristics (Chapter 4).

Applied to real data, I showed that Occ might reveal orders in static sets (Section 5.1). The orders in static sets have to be interpreted based on domain knowledge and might not only give a deeper insight into the relations between classes but also serve as basis for knowledge transfer. In data of developmental stages, the orders representing time steps were revealed in all investigated model organisms, which proved the applicability of Occ for that task (Section 5.2). The application of Occ to cancer data revealed the expected total order in the case of colon carcinoma and subcascades that still support the proposed ordering in case of hepatocellular carcinoma (HCC). However, no consistent order was found in pancreatic adenocarcinoma (PDAC), which underlines the doubt of the order for that particular kind of cancer (Section 5.3).
The exceptional position of PDAC was not only revealed by the Occ analysis but also by analysing the network entropy (Chapter 6). However, a detailed look into the influence of the protein-protein interaction (PPI) on the network entropy rate showed that the results on the course of network entropy across different stages need to be analysed with respect to the underlying PPI.

In conclusion, this thesis provides the theoretical background for defining and distinguishing orders in more than one dimension. I present a technique not only to confirm whether proposed relations between distinct classes are reflected but also to screen for unknown relations in given feature spaces. The application to medically relevant data led to a new hypothesis about relational patterns in leukaemia subgroups and will potentially guide knowledge transfer. Furthermore, my findings support the doubt on the hypothesis of one clear sequence of precursor lesions for PDAC.

LIST OF REFERENCES

[1] Abbas-Aghababazadeh, F., Li, Q., and Fridley, B. L. (2018). Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing. *PLoS ONE*, 13(10).

[2] Abe, S. (2005). *Support Vector Machines for Pattern Classification*. Advances in Computer Vision and Pattern Recognition. Springer-Verlag.

[3] Abreu, R. d. S., Penalva, L. O., Marcotte, E. M., and Vogel, C. (2009). Global signatures of protein and mRNA expression levels. *Molecular bioSystems*, 5(12):1512–1526.

[4] Adamic, L. A. and Adar, E. (2003). Friends and neighbors on the Web. *Social Networks*, 25(3):211–230.

[5] Adhikari, H. and Counter, C. M. (2018). Interrogating the protein interactomes of RAS isoforms identifies PIP5K1A as a KRAS-specific vulnerability. *Nature Communications*, 9(1).

[6] Adler, R. L., Konheim, A. G., and McAndrew, M. H. (1965). Topological entropy. *Transactions of the American Mathematical Society*, 114(2):309–319.

[7] Agba, O. B., Lausser, L., Huse, K., Bergmeier, C., Jahn, N., Groth, M., Bens, M., Sahm, A., Gall, M., Witte, O. W., Kestler, H. A., Schwab, M., and Platzer, M. (2017). Tissue-, sex-, and age-specific DNA methylation of rat glucocorticoid receptor gene promoter and insulin-like growth factor 2 imprinting control region. *Physiological Genomics*, 49(11):690–702.

[8] Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511.

[9] Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169.

[10] Arbeitman, M. N., Furlong, E. E. M., Imam, F., Johnson, E., Null, B. H., Baker, B. S., Krasnow, M. A., Scott, M. P., Davis, R. W., and White, K. P. (2002). Gene expression during the life cycle of Drosophila melanogaster. *Science (New York, N.Y.)*, 297(5590):2270–2275.

[11] Arber, D. A., Orazi, A., Hasserjian, R., Thiele, J., Borowitz, M. J., Le Beau, M. M., Bloomfield, C. D., Cazzola, M., and Vardiman, J. W. (2016). The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood*, 127(20):2391–2405.

[12] Armaghany, T., Wilson, J. D., Chu, Q., and Mills, G. (2012). Genetic alterations in colorectal cancer. *Gastrointestinal cancer research: GCR*, 5(1):19–27.

[13] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

[14] Assou, S., Boumela, I., Haouzi, D., Anahory, T., Dechaud, H., De Vos, J., and Hamamah, S. (2010). Dynamic changes in gene expression during human early embryo development: from fundamental aspects to clinical applications. *Human Reproduction Update*, 17(2):272–290.

[15] Athar, A., Füllgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., Snow, C., Fonseca, N. A., Petryszak, R., Papatheodorou, I., Sarkans, U., and Brazma, A. (2019). ArrayExpress update – from bulk to single-cell expression data. *Nucleic Acids Research*, 47(D1):D711–D715.

[16] Aung, K. L., Fischer, S. E., Denroche, R. E., Jang, G.-H., Dodd, A., Creighton, S., Southwood, B., Liang, S.-B., Chadwick, D., Zhang, A., O'Kane, G. M., Albaba, H., Moura, S., Grant, R. C., Miller, J. K., Mbabaali, F., Pasternack, D., Lungu, I. M., Bartlett, J. M. S., Ghai, S., Lemire, M., Holter, S., Connor, A. A., Moffitt, R. A., Yeh, J. J., Timms, L., Krzyzanowski, P. M., Dhani, N., Hedley, D., Notta, F., Wilson, J. M., Moore, M. J., Gallinger, S., and Knox, J. J. (2018). Genomics-Driven Precision Medicine for Advanced Pancreatic Cancer: Early Results from the COMPASS Trial. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 24(6):1344–1354.

[17] Ayars, M., O'Sullivan, E., Macgregor-Das, A., Shindo, K., Kim, H., Borges, M., Yu, J., Hruban, R. H., Goggins, M., Ayars, M., O'Sullivan, E., Macgregor-Das, A., Shindo, K., Kim, H., Borges, M., Yu, J., Hruban, R. H., and Goggins, M. (2017). IL2RG, identified as overexpressed by RNA-seq profiling of pancreatic intraepithelial neoplasia, mediates pancreatic cancer growth. *Oncotarget*, 8(48):83370–83383.

[18] Banerji, C. R. S., Miranda-Saavedra, D., Severini, S., Widschwendter, M., Enver, T., Zhou, J. X., and Teschendorff, A. E. (2013). Cellular network entropy as the energy potential in Waddington's differentiation landscape. *Scientific Reports*, 3(1).

[19] Banerji, C. R. S., Severini, S., Caldas, C., and Teschendorff, A. E. (2015). Intra-Tumour Signalling Entropy Determines Clinical Outcome in Breast and Lung Cancer. *PLoS Computational Biology*, 11(3).

[20] Barabási, A.-L. and Bonabeau, E. (2003). Scale-Free Networks. *Scientific American*, 288(5):60–69.

[21] Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995.

[22] Basturk, O., Hong, S.-M., Wood, L. D., Adsay, N. V., Albores-Saavedra, J., Biankin, A. V., Brosens, L. A., Fukushima, N., Goggins, M., Hruban, R. H., Kato, Y., Klimstra, D. S., Klöppel, G., Krasinskas, A., Longnecker, D. S., Matthaei, H., Offerhaus, G. J. A., Shimizu, M., Takaori, K., Terris, B., Yachida, S., Esposito, I., and Furukawa, T. (2015). A revised classification system and recommendations from the baltimore consensus meeting for neoplastic precursor lesions in the pancreas. *The American journal of surgical pathology*, 39(12):1730–1741.

[23] Batdorf, B. H., Kroft, S. H., Hosking, P. R., Harrington, A. M., Mackinnon, A. C., and Olteanu, H. (2017). Evaluation of CD43 expression in non-hematopoietic malignancies. *Annals of Diagnostic Pathology*, 29:23–27.

[24] Baugh, L. R., Hill, A. A., Claggett, J. M., Hill-Harfe, K., Wen, J. C., Slonim, D. K., Brown, E. L., and Hunter, C. P. (2005). The homeodomain protein PAL-1 specifies a lineage-specific regulatory network in the C. elegans embryo. *Development (Cambridge, England)*, 132(8):1843–1854.

[25] Bellmann, P. and Schwenker, F. (2020). Ordinal Classification: Working Definition and Detection of Ordinal Structures. *IEEE Access*, 8:164380–164391.

[26] Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

[27] Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

[28] Blanke, C. D. and Faigel, D. O. (2012). 199 - Neoplasms of the Small and Large Intestine. In Goldman, L. and Schafer, A. I., editors, *Goldman's Cecil Medicine (Twenty Fourth Edition)*, pages 1278–1289. W.B. Saunders.

[29] Bork, P., Jensen, L. J., von Mering, C., Ramani, A. K., Lee, I., and Marcotte, E. M. (2004). Protein interaction networks from yeast to human. *Current Opinion in Structural Biology*, 14(3):292–299.

[30] Bost, F. and Kaminski, L. (2019). The metabolic modulator PGC-1α in cancer. *American Journal of Cancer Research*, 9(2):198–211.

[31] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424.

[32] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

[33] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. The Wadsworth statistics/probability series. Chapman and Hall/CRC.

[34] Brenner, H., Hoffmeister, M., Stegmaier, C., Brenner, G., Altenhofen, L., and Haug, U. (2007). Risk of progression of advanced adenomas to colorectal cancer by age and sex: estimates based on 840 149 screening colonoscopies. *Gut*, 56(11):1585–1589.

[35] Brun, C., Chevenet, F., Martin, D., Wojcik, J., Guénoche, A., and Jacq, B. (2004). Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biology*, 5(1):R6.

[36] Buchholz, M., Braun, M., Heidenblut, A., Kestler, H. A., Klöppel, G., Schmiegel, W., Hahn, S. A., Lüttges, J., and Gress, T. M. (2005). Transcriptome analysis of microdissected pancreatic intraepithelial neoplastic lesions. *Oncogene*, 24(44):6626–6636.

[37] Burges, C. J. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

[38] Cardoso, J. S. and Costa, J. F. P. d. (2007). Learning to Classify Ordinal Data: The Data Replication Method. *Journal of Machine Learning Research*, 8:1393–1429.

[39] Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, Ö., Anwar, N., Schultz, N., Bader, G. D., and Sander, C. (2011). Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, 39(Database issue):D685–D690.

[40] Chakraborty, S., Hosen, M. I., Ahmed, M., and Shekhar, H. U. (2018). Onco-Multi-OMICS Approach: A New Frontier in Cancer Research. *BioMed Research International*, 2018.

[41] Chatr-aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., Stark, C., Breitkreutz, B.-J., Dolinski, K., and Tyers, M. (2017). The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, 45(Database issue):D369–D379.

[42] Chen, C. and Lou, T. (2017). Hypoxia inducible factors in hepatocellular carcinoma. *Oncotarget*, 8(28):46691–46703.

[43] Chen, J., Wang, Z., Shen, X., Cui, X., and Guo, Y. (2019a). Identification of novel biomarkers and small molecule drugs in human colorectal cancer by microarray and bioinformatics analysis. *Molecular Genetics & Genomic Medicine*, page e713.

[44] Chen, R., Pan, S., Ottenhof, N. A., de Wilde, R. F., Wolfgang, C. L., Lane, Z., Post, J., Bronner, M. P., Willmann, J. K., Maitra, A., and Brentnall, T. A. (2012). Stromal galectin-1 expression is associated with long-term survival in resectable pancreatic ductal adenocarcinoma. *Cancer Biology & Therapy*, 13(10):899–907.

[45] Chen, W., Jiang, J., Wang, P. P., Gong, L., Chen, J., Du, W., Bi, K., and Diao, H. (2019b). Identifying Hepatocellular Carcinoma Driver Genes by Integrative Pathway Crosstalk and Protein Interaction Network. *DNA and Cell Biology*, 38(10):1112–1124.

[46] Cheng, F., Liu, C., Shen, B., and Zhao, Z. (2016). Investigating cellular network heterogeneity and modularity in cancer: a network entropy and unbalanced motif approach. *BMC Systems Biology*, 10(3):65.

[47] Chou, A., Waddell, N., Cowley, M. J., Gill, A. J., Chang, D. K., Patch, A.-M., Nones, K., Wu, J., Pinese, M., Johns, A. L., Miller, D. K., Kassahn, K. S., Nagrial, A. M., Wasan, H., Goldstein, D., Toon, C. W., Chin, V., Chantrill, L., Humphris, J., Mead, R. S., Rooman, I., Samra, J. S., Pajic, M., Musgrove, E. A., Pearson, J. V., Morey, A. L., Grimmond, S. M., and Biankin, A. V. (2013). Clinical and molecular characterization of HER2 amplified-pancreatic cancer. *Genome Medicine*, 5(8):78.

[48] Chu, L.-F., Leng, N., Zhang, J., Hou, Z., Mamott, D., Vereide, D. T., Choi, J., Kendziorski, C., Stewart, R., and Thomson, J. A. (2016). Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biology*, 17(1).

[49] Colomé-Tatché, M. and Theis, F. J. (2018). Statistical single cell multi-omics integration. *Current Opinion in Systems Biology*, 7:54–59.

[50] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

[51] Cover, T. M. (1965). Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334.

[52] Crick, F. (1970). Central Dogma of Molecular Biology. *Nature*, 227(5258):561–563.

[53] Csardi, G. and Nepusz, T. (2005). The Igraph Software Package for Complex Network Research. *InterJournal*, Complex Systems:1695.

[54] Csete, M. E. and Doyle, J. C. (2002). Reverse Engineering of Biological Complexity. *Science*, 295(5560):1664–1669.

[55] Damjanov, I. (2009). Chapter 5 - Neoplasia. In Damjanov, I., editor, *Pathology Secrets (Third Edition)*, pages 76–97. Mosby.

[56] Demetrius, L. and Manke, T. (2005). Robustness and network evolution—an entropic principle. *Physica A: Statistical Mechanics and its Applications*, 346(3-4):682–696.

[57] Demetrius, L., Matthias Gundlach, V., and Ochs, G. (2004). Complexity and demographic stability in population models. *Theoretical Population Biology*, 65(3):211–225.

[58] Ding, Z. and Kihara, D. (2018). Computational Methods for Predicting Protein-Protein Interactions Using Various Protein Features. *Current protocols in protein science*, 93(1):e62.

[59] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.

[60] Du, Y., Zhao, B., Liu, Z., Ren, X., Zhao, W., Li, Z., You, L., and Zhao, Y. (2017). Molecular Subtyping of Pancreatic Cancer: Translating Genomics and Transcriptomics into the Clinic. *Journal of Cancer*, 8(4):513–522.

[61] Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210.

[62] Emmerich, M. T. M. and Deutz, A. H. (2018). A tutorial on multiobjective optimization: fundamentals and evolutionary methods. *Natural Computing*, 17(3):585–609.

[63] Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A., and Tress, M. L. (2014). Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Human Molecular Genetics*, 23(22):5866–5878.

[64] Fearon, E. R. and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell*, 61(5):759–767.

[65] Feelders, A. (2010). Monotone Relabeling in Ordinal Classification. In *2010 IEEE International Conference on Data Mining*, pages 803–808.

[66] Fix, E. and Hodges, J. (1951). Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties. Technical report, California Univ Berkeley.

[67] Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., and Jensen, L. J. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 41(Database issue):D808–D815.

[68] Frank, E. and Hall, M. A. (2001). A simple approach to ordinal classification. Working Paper, University of Waikato, Department of Computer Science.

[69] Frank, E. and Kramer, S. (2004). Ensembles of nested dichotomies for multi-class problems. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, page 39. Association for Computing Machinery.

[70] Friedell, G. H. (1976). Carcinoma, Carcinoma in Situ, and "Early Lesions" of the Uterine Cervix and the Urinary Bladder: Introduction and Definitions. *Cancer Research*, 36(7 Part 2):2482–2484.

[71] Gan, Y., Li, Y., Li, T., Shu, G., and Yin, G. (2018). CCNA2 acts as a novel biomarker in regulating the growth and apoptosis of colorectal cancer. *Cancer Management and Research*, Volume 10:5113–5124.

[72] Gao, X., Wang, X., and Zhang, S. (2018). Bioinformatics identification of crucial genes and pathways associated with hepatocellular carcinoma. *Bioscience Reports*, 38(6).

[73] Gebauer, F., Tachezy, M., Effenberger, K., von Loga, K., Zander, H., Marx, A., Kaifi, J. T., Sauter, G., Izbicki, J. R., and Bockhorn, M. (2011). Prognostic impact of CXCR4 and CXCR7 expression in pancreatic adenocarcinoma. *Journal of Surgical Oncology*, 104(2):140–145.

[74] Gershgorin, S. (1931). Über die Abgrenzung der Eigenwerte einer Matrix. *Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques et na*, (6):749–754.

[75] Giovannucci, E. (2001). Insulin, Insulin-Like Growth Factors and Colon Cancer: A Review of the Evidence. *The Journal of Nutrition*, 131(11):3109S–3120S.

[76] Gloor, F. J. (1986). The adenoma-carcinoma sequence of the colon and rectum. *Sozial- und Präventivmedizin SPM*, 31(2):74–75.

[77] Gomez-Cabrero, D., Tarazona, S., Ferreirós-Vidal, I., Ramirez, R. N., Company, C., Schmidt, A., Reijmers, T., Paul, V. v. S., Marabita, F., Rodríguez-Ubreva, J., Garcia-Gomez, A., Carroll, T., Cooper, L., Liang, Z., Dharmalingam, G., Kloet, F. v. d., Harms, A. C., Balzano-Nogueira, L., Lagani, V., Tsamardinos, I., Lappe, M., Maier, D., Westerhuis, J. A., Hankemeier, T., Imhof, A., Ballestar, E., Mortazavi, A., Merkenschlager, M., Tegner, J., and Conesa, A. (2019). STATegra, a comprehensive multi-omics dataset of B-cell differentiation in mouse. *Scientific Data*, 6(1):1–15.

[78] Gray, S., Pandha, H. S., Michael, A., Middleton, G., and Morgan, R. (2011). HOX genes in pancreatic development and cancer. *JOP: Journal of the pancreas*, 12(3):216–219.

[79] Grützmann, R., Pilarsky, C., Ammerpohl, O., Lüttges, J., Böhme, A., Sipos, B., Foerder, M., Alldinger, I., Jahnke, B., Schackert, H. K., Kalthoff, H., Kremer, B., Klöppel, G., and Saeger, H. D. (2004). Gene expression profiling of microdissected pancreatic ductal carcinomas using high-density DNA microarrays. *Neoplasia (New York, N.Y.)*, 6(5):611–622.

[80] Gutiérrez, P. A., Pérez-Ortiz, M., Sánchez-Monedero, J., Fernández-Navarro, F., and Hervás-Martínez, C. (2016). Ordinal Regression Methods: Survey and Experimental Study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146.

[81] Hanahan, D. and Weinberg, R. A. (2000). The Hallmarks of Cancer. *Cell*, 100(1):57–70.

[82] Hasan, N. and Ahuja, N. (2019). The Emerging Roles of ATP-Dependent Chromatin Remodeling Complexes in Pancreatic Cancer. *Cancers*, 11(12):1859.

[83] Hashemifar, S., Neyshabur, B., Khan, A. A., and Xu, J. (2018). Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics*, 34(17):i802–i810.

[84] Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The Elements of Statistical Learning*. Springer.

[85] Hawkes, N. (2019). Cancer survival data emphasise importance of early diagnosis. *BMJ*, 364.

[86] He, B., Yin, J., Gong, S., Gu, J., Xiao, J., Shi, W., Ding, W., and He, Y. (2017). Bioinformatics analysis of key genes and pathways for hepatocellular carcinoma transformed from cirrhosis:. *Medicine*, 96(25):e6938.

[87] He, Z., Duan, X., and Zeng, G. (2019). Identification of potential biomarkers and pivotal biological pathways for prostate cancer using bioinformatics analysis methods. *PeerJ*, 7.

[88] Helleputte, T. (2017). *LiblineaR: Linear Predictive Models Based on the LIBLINEAR C/C++ Library*.

[89] Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

[90] Hosoda, W., Chianchiano, P., Griffin, J. F., Pittman, M. E., Brosens, L. A., Noë, M., Yu, J., Shindo, K., Suenaga, M., Rezaee, N., Yonescu, R., Ning, Y., Albores-Saavedra, J., Yoshizawa, N., Harada, K., Yoshizawa, A., Hanada, K., Yonehara, S., Shimizu, M., Uehara, T., Samra, J. S., Gill, A. J., Wolfgang, C. L., Goggins, M. G., Hruban, R. H., and Wood, L. D. (2017). Genetic analyses of isolated high-grade pancreatic intraepithelial neoplasia (HG-PanIN) reveal paucity of alterations in TP53 and SMAD4. *The Journal of Pathology*, 242(1):16–23.

[91] Howlader, N., Noone, A., Krapcho, M., Miller, D., Bishop, K., Kosary, C., Yu, M., Ruhl, J., Tatalovich, Z., Mariotto, A., Lewis, D., Chen, H., Feuer, E., and Cronin, K. (2017). SEER Cancer Statistics Review, 1975-2014. Technical report, National Cancer Institute. Bethesda, MD,.

[92] Hruban, R. H., Adsay, N. V., Albores-Saavedra, J., Compton, C., Garrett, E. S., Goodman, S. N., Kern, S. E., Klimstra, D. S., Klöppel, G., Longnecker, D. S., Lüttges, J., and Offerhaus, G. J. (2001). Pancreatic intraepithelial neoplasia: a new nomenclature and classification system for pancreatic duct lesions. *The American Journal of Surgical Pathology*, 25(5):579–586.

[93] Hruban, R. H., Goggins, M., Parsons, J., and Kern, S. E. (2000). Progression model for pancreatic cancer. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 6(8):2969–2972.

[94] Hu, Q., Qin, Y., Xiang, J., Liu, W., Xu, W., Sun, Q., Ji, S., Liu, J., Zhang, Z., Ni, Q., Xu, J., Yu, X., and Zhang, B. (2018). dCK negatively regulates the NRF2/ARE axis and ROS production in pancreatic cancer. *Cell Proliferation*, 51(4):e12456.

[95] Huhn, J. C. and Hullermeier, E. (2008). Is an ordinal class structure useful in classifier learning? *International Journal of Data Mining, Modelling and Management*, 1(1):45–67.

[96] Ideker, T., Galitski, T., and Hood, L. (2001). A NEW APPROACH TO DECODING LIFE: Systems Biology. *Annual Review of Genomics and Human Genetics*, 2(1):343–372.

[97] International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

[98] Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31(4):e15.

[99] Jaccard, P. (1902). Lois de distribution florale dans la zone alpine. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 38:72.

[100] Jacobson, L. and Sapolsky, R. (1991). The Role of the Hippocampus in Feedback Regulation of the Hypothalamic-Pituitary-Adrenocortical Axis. *Endocrine Reviews*, 12(2):118–134.

[101] Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th Research on Computational Linguistics International Conference*, pages 19–33. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

[102] Jiang, X. and Hao, Y. (2017). Analysis of expression profile data identifies key genes and pathways in hepatocellular carcinoma. *Oncology Letters*.

[103] Johnson, S. K., Dennis, R. A., Barone, G. W., Lamps, L. W., and Haun, R. S. (2006). Differential expression of insulin-like growth factor binding protein-5 in pancreatic adenocarcinomas: identification using DNA microarray. *Molecular Carcinogenesis*, 45(11):814–827.

[104] Ju, C., Colgan, S. P., and Eltzschig, H. K. (2016). Hypoxia-inducible factors as molecular targets for liver diseases. *Journal of Molecular Medicine (Berlin, Germany)*, 94:613–627.
[105] Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30.
[106] Kanwar, N., Nair, R., Wang, D.-Y., and Done, S. J. (2013). Abstract 5116: The calcium channel subunit CACNG4 plays a role in breast cancer metastasis. *Cancer Research*, 73(8 Supplement):5116–5116.
[107] Kauffman, S. A. (1993). *The Origins of Order: Self-organization and Selection in Evolution*. Oxford University Press.
[108] Kendall, M. G. (1955). *Rank Correlation Methods*. C. Griffin.
[109] Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. (2009). Human Protein Reference Database–2009 update. *Nucleic Acids Research*, 37(Database issue):D767–772.
[110] Kitano, H. (2002). Systems biology: a brief overview. *Science (New York, N.Y.)*, 295(5560):1662–1664.
[111] Kohlmann, A., Kipps, T. J., Rassenti, L. Z., Downing, J. R., Shurtleff, S. A., Mills, K. I., Gilkes, A. F., Hofmann, W.-K., Basso, G., Dell'Orto, M. C., Foà, R., Chiaretti, S., De Vos, J., Rauhut, S., Papenhausen, P. R., Hernández, J. M., Lumbreras, E., Yeoh, A. E., Koay, E. S., Li, R., Liu, W.-m., Williams, P. M., Wieczorek, L., and Haferlach, T. (2008). An international standardization programme towards the application of gene expression profiling in routine leukaemia diagnostics: the Microarray Innovations in LEukemia study prephase. *British Journal of Haematology*, 142(5):802–807.
[112] Kohonen, T. (1995). *Self-Organizing Maps*, volume I. Springer-Verlag.
[113] Korbel, J. O. and Campbell, P. J. (2013). Criteria for Inference of Chromothripsis in Cancer Genomes. *Cell*, 152(6):1226–1236.
[114] Kozuka, S., Sassa, R., Taki, T., Masamoto, K., Nagasawa, S., Saga, S., Hasegawa, K., and Takeuchi, M. (1979). Relation of pancreatic duct hyperplasia to carcinoma. *Cancer*, 43(4):1418–1428.
[115] Kraus, J. M. and Kestler, H. A. (2010). A highly efficient multi-core algorithm for clustering extremely large datasets. *BMC Bioinformatics*, 11(1):169.
[116] Kuchaiev, O., Rasajski, M., Higham, D. J., and Przulj, N. (2009). Geometric de-noising of protein-protein interaction networks. *PLoS computational biology*, 5(8):e1000454.
[117] Kuo, T.-L., Cheng, K.-H., Chen, L.-T., and Hung, W.-C. (2019). Deciphering The Potential Role of Hox Genes in Pancreatic Cancer. *Cancers*, 11(5):734.
[118] Lascorz, J., Chen, B., Hemminki, K., and Försti, A. (2011). Consensus Pathways Implicated in Prognosis of Colorectal Cancer Identified Through Systematic Enrichment Analysis of Gene Expression Profiling Studies. *PLoS ONE*, 6(4):e18867.
[119] Lattke*, R., Lausser*, L., Müssel, C., and Kestler, H. A. (2015). Detecting Ordinal Class Structures. In *Multiple Classifier Systems*, Lecture Notes in Computer Science, pages 100–111. Springer, Cham (*equal contribution).
[120] Lausser*, L., Schäfer*, L. M., and Kestler, H. A. (2018). Ordinal classifiers can fail on repetitive class structures. *Archives of Data Science, Series A*, (*equal contribution).
[121] Lausser*, L., Schäfer*, L. M., and Kestler, H. A. (2020a). ORION: Ordinal Relations. *R package version 1.0.*, (*equal contribution).
[122] Lausser*, L., Schäfer*, L. M., Kühlwein, S. D., Kestler, A. M., and Kestler, H. A. (2020b). Detecting ordinal subcascades. *Neural Processing Letters*, (*equal contribution).
[123] Lausser*, L., Schäfer*, L. M., Schirra*, L.-R., Szekely, R., Schmid, F., and Kestler, H. A. (2019). Assessing phenotype order in molecular data. *Scientific Reports*, 9(1):1–10 (*equal contribution).
[124] Leon-Mimila, P., Wang, J., and Huertas-Vazquez, A. (2019). Relevance of Multi-Omics Studies in Cardiovascular Diseases. *Frontiers in Cardiovascular Medicine*, 6.
[125] Li, F., Li, Q., and Wu, X. (2018). Construction and analysis for differentially expressed long non-coding RNAs and MicroRNAs mediated competing endogenous RNA network in colon cancer. *PLoS ONE*, 13(2).
[126] Li, J. J., Bickel, P. J., and Biggin, M. D. (2014). System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ*, 2.
[127] Li, J. J. and Biggin, M. D. (2015). Statistics requantitates the central dogma. *Science*, 347(6226):1066–1067.
[128] Li, Y., Chen, X., Huang, H., Li, G., Liao, L., Yuan, T., and Deng, S. (2019). Identification Of Novel lncRNAs For Detection Of HBV-Associated Hepatocellular Carcinoma. *OncoTargets and Therapy*, Volume 12:10199–10211.
[129] Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.
[130] Lin, D. (1998). An Information-Theoretic Definition of Similarity. In *In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann.
[131] Liu, F., Ji, F., Ji, Y., Jiang, Y., Sun, X., Lu, Y., Zhang, L., Han, Y., and Liu, X. (2015). In-depth analysis of the critical genes and pathways in colorectal cancer. *International Journal of Molecular Medicine*, 36(4):923–930.
[132] Liu, Y., Beyer, A., and Aebersold, R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell*, 165(3):535–550.
[133] Liu, Y.-C., Yeh, C.-T., and Lin, K.-H. (2019a). Molecular Functions of Thyroid Hormone Signaling in Regulation of Cancer Progression and Anti-Apoptosis. *International Journal of Molecular Sciences*, 20(20).
[134] Liu, Y.-R., Hu, Y., Zeng, Y., Li, Z.-X., Zhang, H.-B., Deng, J.-L., and Wang, G. (2019b). Neurexophilin and PC-esterase domain family member 4 (NXPE4) and prostate androgen-regulated mucin-like protein 1 (PARM1) as prognostic biomark-

ers for colorectal cancer. *Journal of Cellular Biochemistry*, 120(10):18041–18052.
[135] Loeb, K. R. and Loeb, L. A. (2000). Significance of multiple mutations in cancer. *Carcinogenesis*, 21(3):379–385.
[136] Lovén, J., Orlando, D. A., Sigova, A. A., Lin, C. Y., Rahl, P. B., Burge, C. B., Levens, D. L., Lee, T. I., and Young, R. A. (2012). Revisiting Global Gene Expression Analysis. *Cell*, 151(3):476–482.
[137] Luijten, M. N. H., Lee, J. X. T., and Crasta, K. C. (2018). Mutational game changer: Chromothripsis and its emerging relevance to cancer. *Mutation Research/Reviews in Mutation Research*, 777:29–51.
[138] Lv, J. and Li, L. (2019). Hub Genes and Key Pathway Identification in Colorectal Cancer Based on Bioinformatic Analysis. *BioMed Research International*, 2019:1–13.
[139] Ma, X., Sun, H., Fu, J., Gao, X., and Li, X. (2016). Identification of potential biomarkers in pathogenesis and liver metastasis of colorectal cancer: a bioinformatics study. *International Journal of Clinical and Experimental Medicine*, 9(12):23198–23205.
[140] Ma, X., Zhou, L., and Zheng, S. (2020). Transcriptome analysis revealed key prognostic genes and microRNAs in hepatocellular carcinoma. *PeerJ*, 8.
[141] Mahdavi, M. A. and Lin, Y.-H. (2007). False positive reduction in protein-protein interaction predictions using gene ontology annotations. *BMC Bioinformatics*, 8(1):262.
[142] Mahmud, I. and Liao, D. (2019). DAXX in cancer: phenomena, processes, mechanisms and regulation. *Nucleic Acids Research*, 47(15):7734–7752.
[143] Maitra, A., Hansel, D. E., Argani, P., Ashfaq, R., Rahman, A., Naji, A., Deng, S., Geradts, J., Hawthorne, L., House, M. G., and Yeo, C. J. (2003). Global expression analysis of well-differentiated pancreatic endocrine neoplasms using oligonucleotide microarrays. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 9(16 Pt 1):5988–5995.
[144] Malone, J. H. and Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biology*, 9(1).
[145] Maltese, G. (1986). A Simple Proof of the Fundamental Theorem of Finite Markov Chains. *The American Mathematical Monthly*, 93(8):629.
[146] Marbach, D., Schaffter, T., Floreano, D., Prill, R. J., and Stolovitzky, G. (2009). The dream4 in-silico network challenge. *Draft, version 0.3*.
[147] Marotta, M., Chen, X., Inoshita, A., Stephens, R., Thomas Budd, G., Crowe, J. P., Lyons, J., Kondratova, A., Tubbs, R., and Tanaka, H. (2012). A common copy-number breakpoint of ERBB2 amplification in breast cancer colocalizes with a complex block of segmental duplications. *Breast Cancer Research : BCR*, 14(6):R150.
[148] Marquardt, J. U., Andersen, J. B., and Thorgeirsson, S. S. (2015). Functional and genetic deconstruction of the cellular origin in liver cancer. *Nature Reviews Cancer*, 15(11):653–667.
[149] McInnes, E. (2017). General Pathology and the Terminology of Basic Pathology. In *Pathology of Toxicologists*, pages 35–58. John Wiley & Sons, Ltd.
[150] Melssen, W., Wehrens, R., and Buydens, L. (2006). Supervised Kohonen networks for classification problems. *Chemometrics and Intelligent Laboratory Systems*, 83(2):99–113.
[151] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2019). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*.
[152] Michalski, C. W., Erkan, M., Sauliunaite, D., Giese, T., Stratmann, R., Sartori, C., Giese, N. A., Friess, H., and Kleeff, J. (2008). Ex vivo chemosensitivity testing and gene expression profiling predict response towards adjuvant gemcitabine treatment in pancreatic cancer. *British Journal of Cancer*, 99(5):760–767.
[153] Mieczkowski, J., Swiatek-Machado, K., and Kaminska, B. (2012). Identification of Pathway Deregulation – Gene Expression Based Analysis of Consistent Signal Transduction. *PLoS ONE*, 7(7).
[154] Miller, G. A., Miller, G., and Miller, G. A. (1955). Note on the bias of information estimates. *Information Theory in Psychology: Problems and Methods*.
[155] Moris, N., Pina, C., and Arias, A. M. (2016). Transition states and cell fate decisions in epigenetic landscapes. *Nature Reviews Genetics*, 17(11):693–703.
[156] Morson, B. (1974). The Polyp-cancer Sequence in the Large Bowel. *Proceedings of the Royal Society of Medicine*, 67(6 Pt 1):451–457.
[157] Murphy, S. J., Hart, S. N., Lima, J. F., Kipp, B. R., Klebig, M., Winters, J. L., Szabo, C., Zhang, L., Eckloff, B. W., Petersen, G. M., Scherer, S. E., Gibbs, R. A., McWilliams, R. R., Vasmatzis, G., and Couch, F. J. (2013). Genetic Alterations Associated With Progression From Pancreatic Intraepithelial Neoplasia to Invasive Pancreatic Tumor. *Gastroenterology*, 145(5):1098–1109.e1.
[158] Müssel*, C., Lausser*, L., Maucher, M., and Kestler, H. A. (2012). Multi-Objective Parameter Selection for Classifiers. *Journal of Statistical Software*, 46(1):1–27 (*equal contribution).
[159] Newman, S. A. (2006). The Developmental Genetic Toolkit and the Molecular Homology—Analogy Paradox. *Biological Theory*, 1(1):12–16.
[160] Nilsson, R., Peña, J. M., Björkegren, J., and Tegnér, J. (2006). Evaluating Feature Selection for SVMs in High Dimensions. In Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Fürnkranz, J., Scheffer, T., and Spiliopoulou, M., editors, *Machine Learning: ECML 2006*, volume 4212, pages 719–726. Springer Berlin Heidelberg.

[161] Notta, F., Chan-Seng-Yue, M., Lemire, M., Li, Y., Wilson, G. W., Connor, A. A., Denroche, R. E., Liang, S.-B., Brown, A. M., Kim, J. C., Wang, T., Simpson, J. T., Beck, T., Borgida, A., Buchner, N., Chadwick, D., Hafezi-Bakhtiari, S., Dick, J. E., Heisler, L., Hollingsworth, M. A., Ibrahimov, E., Jang, G. H., Johns, J., Jorgensen, L. G., Law, C., Ludkovski, O., Lungu, I., Ng, K., Pasternack, D., Petersen, G. M., Shlush, L. I., Timms, L., Tsao, M.-S., Wilson, J. M., Yung, C. K., Zogopoulos, G., Bartlett, J. M., Alexandrov, L. B., Real, F. X., Cleary, S. P., Roehrl, M. H., McPherson, J. D., Stein, L. D., Hudson, T. J., Campbell, P. J., and Gallinger, S. (2016). A renewed model of pancreas cancer evolution based on genomic rearrangement patterns. *Nature*, 538(7625):378–382.

[162] Notta, F., Hahn, S. A., and Real, F. X. (2017). A genetic roadmap of pancreatic cancer: still evolving. *Gut*, 66(12):2170–2178.

[163] Ohhashi, S., Ohuchida, K., Mizumoto, K., Egami, T., Yu, J., Cui, L., Toma, H., Takahata, S., Nabae, T., and Tanaka, M. (2009). Midkine mRNA is overexpressed in pancreatic cancer. *Digestive Diseases and Sciences*, 54(4):811–815.

[164] Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., Kolas, N., O'Donnell, L., Leung, G., McAdam, R., Zhang, F., Dolma, S., Willems, A., Coulombe-Huntington, J., Chatr-aryamontri, A., Dolinski, K., and Tyers, M. (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Research*, 47(D1):D529–D541.

[165] Pan, S., Chen, R., Stevens, T., Bronner, M. P., May, D., Tamura, Y., McIntosh, M. W., and Brentnall, T. A. (2011). Proteomics portrait of archival lesions of chronic pancreatitis. *PloS One*, 6(11):e27574.

[166] Paninski, L. (2003). Estimation of Entropy and Mutual Information. *Neural Computation*, 15(6):1191–1253.

[167] Parsa, N. (2012). Environmental Factors Inducing Human Cancers. *Iranian Journal of Public Health*, 41(11):1–9.

[168] Pearl, J. (2003). Causality: models, reasoning and inference. *Econometric Theory*, 19(675-685):46.

[169] Pearson, K. and Galton, F. (1895). VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242.

[170] Pelosi, E., Castelli, G., and Testa, U. (2017). Pancreatic Cancer: Molecular Characterization, Clonal Evolution and Cancer Stem Cells. *Biomedicines*, 5(4).

[171] Penque, D. (2009). Two-dimensional gel electrophoresis and mass spectrometry for biomarker discovery. *PROTEOMICS – Clinical Applications*, 3(2):155–172.

[172] Pira, G., Uva, P., Scanu, A. M., Rocca, P. C., Murgia, L., Uleri, E., Piu, C., Porcu, A., Carru, C., Manca, A., Persico, I., Muroni, M. R., Sanges, F., Serra, C., Dolei, A., Angius, A., and De Miglio, M. R. (2020). Landscape of transcriptome variations uncovering known and novel driver events in colorectal carcinoma. *Scientific Reports*, 10(1).

[173] Platt, J. C., Cristianini, N., and Shawe-Taylor, J. (1999). Large margin DAGs for multiclass classification. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, pages 547–553. MIT Press.

[174] Polvani, S. (2014). Nuclear receptors and pathogenesis of pancreatic cancer. *World Journal of Gastroenterology*, 20(34):12062.

[175] Prasad, N. B., Biankin, A. V., Fukushima, N., Maitra, A., Dhara, S., Elkahloun, A. G., Hruban, R. H., Goggins, M., and Leach, S. D. (2005). Gene expression profiles in pancreatic intraepithelial neoplasia reflect the effects of Hedgehog signaling on pancreatic ductal epithelial cells. *Cancer Research*, 65(5):1619–1626.

[176] Qu, X., Sandmann, T., Frierson, H., Fu, L., Fuentes, E., Walter, K., Okrah, K., Rumpel, C., Moskaluk, C., Lu, S., Wang, Y., Bourgon, R., Penuel, E., Pirzkall, A., Amler, L., Lackner, M. R., Tabernero, J., Hampton, G. M., and Kabbarah, O. (2016). Integrated genomic analysis of colorectal cancer progression reveals activation of EGFR through demethylation of the EREG promoter. *Oncogene*, 35(50):6403–6415.

[177] Rajmohan, V. and Mohandas, E. (2007). The limbic system. *Indian Journal of Psychiatry*, 49(2):132–139.

[178] Ramanan, V. K., Shen, L., Moore, J. H., and Saykin, A. J. (2012). Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends in Genetics*, 28(7):323–332.

[179] Rawla, P., Sunkara, T., and Barsouk, A. (2019). Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. *Przeglad Gastroenterologiczny*, 14(2):89–103.

[180] Rawnaq, T., Dietrich, L., Wolters-Eisfeld, G., Uzunoglu, F. G., Vashist, Y. K., Bachmann, K., Simon, R., Izbicki, J. R., Bockhorn, M., and Güngör, C. (2014). The Multifunctional Growth Factor Midkine Promotes Proliferation and Migration in Pancreatic Cancer. *Molecular Cancer Research*, 12(5):670–680.

[181] Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359.

[182] Real, F. X. (2003). A "catastrophic hypothesis" for pancreas cancer progression. *Gastroenterology*, 124(7):1958–1964.

[183] Real, F. X. and Andrés, M. P. d. (2018). Mutant Kras Dosage and Chromothripsis: The Right Ingredients for a Pancreatic Cancer Catastrophe. *Trends in Cancer*, 4(6):399–401.

[184] Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *arXiv:cmp-lg/9511007*.

[185] Sato, M., Kawana, K., Adachi, K., Fujimoto, A., Yoshida, M., Nakamura, H., Nishida, H., Inoue, T., Taguchi, A., Ogishima, J., Eguchi, S., Yamashita, A., Tomio, K., Wada-Hiraike, O., Oda, K., Nagamatsu, T., Osuga, Y., and Fujii, T. (2017). Intracellular signaling entropy can be a biomarker for predicting the development of cervical intraepithelial neoplasia. *PLoS ONE*, 12(4).

[186] Schaffter, T., Marbach, D., and Floreano, D. (2011). GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270.

[187] Schlicker, A., Domingues, F. S., Rahnenführer, J., and Lengauer, T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, 7(1):302.

[188] Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., Fulton, R. S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K. M., Auger, K., Chow, W., Collins, J., Harden, G., Hubbard, T., Pelan, S., Simpson, J. T., Threadgold, G., Torrance, J., Wood, J. M., Clarke, L., Koren, S., Boitano, M., Peluso, P., Li, H., Chin, C.-S., Phillippy, A. M., Durbin, R., Wilson, R. K., Flicek, P., Eichler, E. E., and Church, D. M. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, 27(5):849–864.

[189] Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342.

[190] Schöning, U. and Kestler, H. A. (2016). *Mathe-Toolbox*. Lehmanns Media, 3. edition.

[191] Shangguan, H., Tan, S.-Y., and Zhang, J.-R. (2015). Bioinformatics analysis of gene expression profiles in hepatocellular carcinoma. *European Review for Medical and Pharmacological Sciences*, 19(11):2054–2061.

[192] Sheffer, M., Bacolod, M. D., Zuk, O., Giardina, S. F., Pincas, H., Barany, F., Paty, P. B., Gerald, W. L., Notterman, D. A., and Domany, E. (2009). Association of survival and disease progression with chromosomal instability: A genomic exploration of colorectal cancer. *Proceedings of the National Academy of Sciences*, 106(17):7131–7136.

[193] Shen, Q., Eun, J. W., Lee, K., Kim, H. S., Yang, H. D., Kim, S. Y., Lee, E. K., Kim, T., Kang, K., Kim, S., Min, D.-H., Oh, S.-N., Lee, Y.-J., Moon, H., Ro, S. W., Park, W. S., Lee, J. Y., and Nam, S. W. (2018). Barrier to autointegration factor 1, procollagen-lysine, 2-oxoglutarate 5-dioxygenase 3, and splicing factor 3b subunit 4 as early-stage cancer decision markers and drivers of hepatocellular carcinoma. *Hepatology (Baltimore, Md.)*, 67(4):1360–1377.

[194] Shibata, W., Kinoshita, H., Hikiba, Y., Sato, T., Ishii, Y., Sue, S., Sugimori, M., Suzuki, N., Sakitani, K., Ijichi, H., Mori, R., Endo, I., and Maeda, S. (2018). Overexpression of HER2 in the pancreas promotes development of intraductal papillary mucinous neoplasms in mice. *Scientific Reports*, 8.

[195] Smith, S. M. and Vale, W. W. (2006). The role of the hypothalamic-pituitary-adrenal axis in neuroendocrine responses to stress. *Dialogues in Clinical Neuroscience*, 8(4):383–395.

[196] Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1):72–101.

[197] Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science*, 302(5643):249–255.

[198] Stumpf, M. P. H., Thorne, T., Silva, E. d., Stewart, R., An, H. J., Lappe, M., and Wiuf, C. (2008). Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences*, 105(19):6959–6964.

[199] Sun, S., Li, Y., Han, S., Jia, H., Li, X., and Li, X. (2019). A comprehensive genome-wide profiling comparison between HBV and HCV infected hepatocellular carcinoma. *BMC Medical Genomics*, 12(1).

[200] Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J., and von Mering, C. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(Database issue):D447–452.

[201] Särndal, C. E. (1974). A comparative study of association measures. *Psychometrika*, 39(2):165–187.

[202] Tao, C., Sun, J., Zheng, W. J., Chen, J., and Xu, H. (2015). Colorectal cancer drug target prediction using ontology-based inference and network analysis. *Database*, 2015.

[203] Tao, Y., Tian, K., Chen, J., Tan, D., Liu, Y., Xiong, Y., Chen, Z., and Tian, Y. (2019). Network Pharmacology-Based Prediction of the Active Compounds, Potential Targets, and Signaling Pathways Involved in Danshiliuhao Granule for Treatment of Liver Fibrosis. *Evidence-based Complementary and Alternative Medicine : eCAM*, 2019.

[204] Teschendorff, A. E., Banerji, C. R. S., Severini, S., Kuehn, R., and Sollich, P. (2015). Increased signaling entropy in cancer requires the scale-free property of proteininteraction networks. *Scientific Reports*, 5(1):9646.

[205] Teschendorff, A. E. and Severini, S. (2010). Increased entropy of signal transduction in the cancer metastasis phenotype. *BMC Systems Biology*, 4:104.

[206] Thanki, K., Nicholls, M. E., Gajjar, A., Senagore, A. J., Qiu, S., Szabo, C., Hellmich, M. R., and Chao, C. (2017). Consensus Molecular Subtypes of Colorectal Cancer and their Clinical Implications. *International biological and biomedical journal*, 3(3):105–111.

[207] The Gene Ontology Consortium (2017). Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research*, 45(D1):D331–D338.

[208] The Gene Ontology Consortium (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338.

[209] Thul, P. J. and Lindskog, C. (2018). The human protein atlas: A spatial map of the human proteome. *Protein Science : A Publication of the Protein Society*, 27(1):233–244.

[210] Toyama, R., Chen, X., Jhawar, N., Aamar, E., Epstein, J., Reany, N., Alon, S., Gothilf, Y., Klein, D. C., and Dawid, I. B. (2009). Transcriptome analysis of the zebrafish pineal gland. *Developmental Dynamics: An Official Publication of the American Association of Anatomists*, 238(7):1813–1826.

[211] Toyoda, E., Doi, R., Kami, K., Mori, T., Ito, D., Koizumi, M., Kida, A., Nagai, K., Ito, T., Masui, T., Wada, M., Tagawa, M., and Uemoto, S. (2008). Midkine promoter-based conditionally replicative adenovirus therapy for midkine-expressing human pancreatic cancer. *Journal of experimental & clinical cancer research: CR*, 27:30.

[212] Tukey, J. W. (1975). Mathematics and the Picturing of Data. *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, 2:523–531.

[213] Ulgen, E., Ozisik, O., and Sezerman, O. U. (2019). pathfindR: An R Package for Comprehensive Identification of Enriched Pathways in Omics Data Through Active Subnetworks. *Frontiers in Genetics*, 10.

[214] Vapnik, V. N. (1998). *Statistical learning theory*. Adaptive and learning systems for signal processing, communications, and control. Wiley.

[215] Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, fourth edition.

[216] Vincent, A., Herman, J., Schulick, R., Hruban, R. H., and Goggins, M. (2011). Pancreatic cancer. *The Lancet*, 378(9791):607–620.

[217] Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I.

[218] Vogel, C. and Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews. Genetics*, 13(4):227–232.

[219] von Mering, C. (2004). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33(Database issue):D433–D437.

[220] Waddell, N., Pajic, M., Patch, A.-M., Chang, D. K., Kassahn, K. S., Bailey, P., Johns, A. L., Miller, D., Nones, K., Quek, K., Quinn, M. C. J., Robertson, A. J., Fadlullah, M. Z. H., Bruxner, T. J. C., Christ, A. N., Harliwong, I., Idrisoglu, S., Manning, S., Nourse, C., Nourbakhsh, E., Wani, S., Wilson, P. J., Markham, E., Cloonan, N., Anderson, M. J., Fink, J. L., Holmes, O., Kazakoff, S. H., Leonard, C., Newell, F., Poudel, B., Song, S., Taylor, D., Waddell, N., Wood, S., Xu, Q., Wu, J., Pinese, M., Cowley, M. J., Lee, H. C., Jones, M. D., Nagrial, A. M., Humphris, J., Chantrill, L. A., Chin, V., Steinmann, A. M., Mawson, A., Humphrey, E. S., Colvin, E. K., Chou, A., Scarlett, C. J., Pinho, A. V., Giry-Laterriere, M., Rooman, I., Samra, J. S., Kench, J. G., Pettitt, J. A., Merrett, N. D., Toon, C., Epari, K., Nguyen, N. Q., Barbour, A., Zeps, N., Jamieson, N. B., Graham, J. S., Niclou, S. P., Bjerkvig, R., Grützmann, R., Aust, D., Hruban, R. H., Maitra, A., Iacobuzio-Donahue, C. A., Wolfgang, C. L., Morgan, R. A., Lawlor, R. T., Corbo, V., Bassi, C., Falconi, M., Zamboni, G., Tortora, G., Tempero, M. A., Gill, A. J., Eshleman, J. R., Pilarsky, C., Scarpa, A., Musgrove, E. A., Pearson, J. V., Biankin, A. V., and Grimmond, S. M. (2015). Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature*, 518(7540):495–501.

[221] Waddington, C. H. (1957). The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser. *The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser.*

[222] Wan, Z., Zhang, X., Luo, Y., and Zhao, B. (2019). Identification of Hepatocellular Carcinoma-Related Potential Genes and Pathways Through Bioinformatic-Based Analyses. *Genetic Testing and Molecular Biomarkers*, 23(11):766–777.

[223] Wang, B.-s., Liu, Z., Sun, S.-l., and Zhao, Y. (2014). Identification of genes and candidate agents associated with pancreatic cancer. *Tumor Biology*, 35(1):81–88.

[224] Wang, D., Eraslan, B., Wieland, T., Hallström, B., Hopf, T., Zolg, D. P., Zecha, J., Asplund, A., Li, L.-h., Meng, C., Frejno, M., Schmidt, T., Schnatbaum, K., Wilhelm, M., Ponten, F., Uhlen, M., Gagneur, J., Hahne, H., and Kuster, B. (2019). A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Molecular Systems Biology*, 15(2):e8503.

[225] Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., and Chen, C.-F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23(10):1274–1281.

[226] Wehrens, R. and Buydens, L. M. C. (2007). Self- and Super-organizing Maps in R: The kohonen Package. *Journal of Statistical Software*, 21(1):1–19.

[227] Wehrens, R. and Kruisselbrink, J. (2018). Flexible Self-Organizing Maps in kohonen 3.0. *Journal of Statistical Software*, 87(1):1–18.

[228] Weinberg, R. (2013). *The Biology of Cancer, Second Edition*. Garland Science.

[229] Wen, F., Shen, A., Choi, A., Gerner, E. W., and Shi, J. (2013). Extracellular DNA in Pancreatic Cancer Promotes Cell Invasion and Metastasis. *Cancer Research*, 73(14):4256–4266.

[230] West, J., Bianconi, G., Severini, S., and Teschendorff, A. E. (2012). Differential network entropy reveals cancer system hallmarks. *Scientific Reports*, 2(1).

[231] Whitacre, J. M. (2012). Biological Robustness: Paradigms, Mechanisms, and Systems Principles. *Frontiers in Genetics*, 3.

[232] Wiernik, P. H., Dutcher, J. P., and Gertz, M. A. (2018). *Neoplastic Diseases of the Blood*. Springer.

[233] Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann series in data management systems. Morgan Kaufmann, 3rd ed edition.

[234] Wu, M., Liu, Z., Zhang, A., and Li, N. (2019). Identification of key genes and pathways in hepatocellular carcinoma: A preliminary bioinformatics analysis. *Medicine*, 98(5):e14287.

[235] Wu, Y., Krueger, G. R. F., and Wang, G. (2016). Altered Micro-RNA Degradation Promotes Tumor Heterogeneity: A Result from Boolean Network Modeling. *Anticancer Research*, 36(2):575–581.

[236] Xiao, Y., Yurievich, U. A., and Yosypovych, S. V. (2017). Long noncoding RNA XIST is a prognostic factor in colorectal cancer and inhibits 5-fluorouracil-induced cell cytotoxicity through promoting thymidylate synthase expression. *Oncotarget*, 8(47):83171–83182.

[237] Xie, S., Jiang, X., Zhang, J., Xie, S., Hua, Y., Wang, R., and Yang, Y. (2019). Identification of significant gene and pathways involved in HBV-related hepatocellular carcinoma by bioinformatics analysis. *PeerJ*, 7:e7408.

[238] Xiong, J., Altaf, K., Ke, N., Wang, Y., Tang, J., Tan, C., Li, A., Zhang, H., He, D., and Liu, X. (2016). dCK Expression and Gene Polymorphism With Gemcitabine Chemosensitivity in Patients With Pancreatic Ductal Adenocarcinoma: A

Strobe-Compliant Observational Study. *Medicine*, 95(10):e2936.

[239] Yaeger, R., Chatila, W. K., Lipsyc, M. D., Hechtman, J. F., Cercek, A., Sanchez-Vega, F., Jayakumaran, G., Middha, S., Zehir, A., Donoghue, M. T., You, D., Viale, A., Kemeny, N., Segal, N. H., Stadler, Z. K., Varghese, A. M., Kundra, R., Gao, J., Syed, A., Hyman, D. M., Vakiani, E., Rosen, N., Taylor, B. S., Ladanyi, M., Berger, M. F., Solit, D. B., Shia, J., Saltz, L., and Schultz, N. (2018). Clinical sequencing defines the genomic landscape of metastatic colorectal cancer. *Cancer cell*, 33(1):125–136.e3.

[240] Yang, C., Ren, J., Li, B., Jin, C., Ma, C., Cheng, C., Sun, Y., and Shi, X. (2019). Identification of gene biomarkers in patients with postmenopausal osteoporosis. *Molecular Medicine Reports*, 19(2):1065–1073.

[241] Yassin, M., Kissow, H., Vainer, B., Joseph, P. D., Hay-Schmidt, A., Olsen, J., and Pedersen, A. E. (2018). Cytoglobin affects tumorigenesis and the expression of ulcerative colitis-associated genes under chemically induced colitis in mice. *Scientific Reports*, 8(1):1–16.

[242] Yin, S., Dou, J., Yang, G., and Chen, F. (2019). Long non-coding RNA XIST expression as a prognostic factor in human cancers: A meta-analysis. *The International Journal of Biological Markers*, 34(4):327–333.

[243] Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7):976–978.

[244] Yu, L., Xu, F., and Gao, L. (2017). Predicting therapeutic and aggravating drugs for hepatocellular carcinoma based on tissue-specific pathways. *arXiv:1707.00954 [q-bio]*.

[245] Yu, M.-H., Luo, Y., Qin, S.-L., Wang, Z.-S., Mu, Y.-F., and Zhong, M. (2015). Up-regulated CKS2 promotes tumor progression and predicts a poor prognosis in human colorectal cancer. *American Journal of Cancer Research*, 5(9):2708–2718.

[246] Yuan, C., Bao, Y., Sato, K., Nimptsch, K., Song, M., Brand-Miller, J. C., Morales-Oyarvide, V., Zoltick, E. S., Keum, N., Wolpin, B. M., Meyerhardt, J. A., Chan, A. T., Willett, W. C., Stampfer, M. J., Wu, K., Giovannucci, E. L., Fuchs, C. S., and Ng, K. (2017). Influence of dietary insulin scores on survival in colorectal cancer patients. *British Journal of Cancer*, 117(7):1079–1087.

[247] Zahid, K. R., Yao, S., Khan, A. R. R., Raza, U., and Gou, D. (2019). mTOR/HDAC1 Crosstalk Mediated Suppression of ADH1A and ALDH2 Links Alcohol Metabolism to Hepatocellular Carcinoma Onset and Progression in silico. *Frontiers in Oncology*, 9.

[248] Zhang, C., Peng, L., Zhang, Y., Liu, Z., Li, W., Chen, S., and Li, G. (2017). The identification of key genes and pathways in hepatocellular carcinoma by bioinformatics analysis of high-throughput data. *Medical Oncology*, 34(6).

[249] Zhang, G., He, P., Tan, H., Budhu, A., Gaedcke, J., Ghadimi, B. M., Ried, T., Yfantis, H. G., Lee, D. H., Maitra, A., Hanna, N., Alexander, H. R., and Hussain, S. P. (2013). Integration of Metabolomics and Transcriptomics Revealed a Fatty Acid Network Exerting Growth Inhibitory Effects in Human Pancreatic Cancer. *Clinical Cancer Research*, 19(18):4983–4993.

[250] Zhang, G.-L., Pan, L.-L., Huang, T., and Wang, J.-H. (2019). The transcriptome difference between colorectal tumor and normal tissues revealed by single-cell sequencing. *Journal of Cancer*, 10(23):5883–5890.

[251] Zhang, J., Roebuck, P., and Coombes, K. (2012a). Simulating gene expression data to estimate sample size for class and biomarker discovery. *Int J Advances Life Sci*, 4:44–51.

[252] Zhang, L., Song, X., Shao, Y., Wu, C., and Jiang, J. (2018). Prognostic value of Midkine expression in patients with solid tumors: a systematic review and meta-analysis. *Oncotarget*, 9(37).

[253] Zhang, X., Zhang, M., Hou, Y., Xu, L., Li, W., Zou, Z., Liu, C., Xu, A., and Wu, S. (2016). Single-cell analyses of transcriptional heterogeneity in squamous cell carcinoma of urinary bladder. *Oncotarget*, 7(40):66069–66076.

[254] Zhang, X., Zhao, X.-M., He, K., Lu, L., Cao, Y., Liu, J., Hao, J.-K., Liu, Z.-P., and Chen, L. (2012b). Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics*, 28(1):98–104.

[255] Zhao, Z., Fan, X., Yang, L., Song, J., Fang, S., Tu, J., Zheng, L., Wu, F., Zhang, D., Ying, X., and Ji, J. (2019). The identification of a common different gene expression signature in patients with colorectal cancer. *Mathematical Biosciences and Engineering*, 16(4):17.

[256] Zhu, H., Ji, Y., Li, W., and Wu, M. (2019). Identification of key pathways and genes in colorectal cancer to predict the prognosis based on mRNA interaction network. *Oncology Letters*.

[257] Zhu, S., Yu, W., Yang, X., Wu, C., and Cheng, F. (2020). Traditional Classification and Novel Subtyping Systems for Bladder Cancer. *Frontiers in Oncology*, 10.

APPENDIX

Ordinal classifier cascade - Preselection strategy (Section 3.1.1)

The following provides the proof of Theorem 1. This proof was published in the Supplementary to Lausser* et al. [123]. It is based on Lemmata A1 and A2, which are derived from the more general Theorems A1 and A2.

Theorem A1. Let h denote a full ordinal classifier cascade

$$h : \mathbb{R}^n \longrightarrow y = \left\{ y_{(1)}, \ldots, y_{(|y|)} \right\}. \tag{A1}$$

Let furthermore i,j,l denote indices with $1 \leqslant i \leqslant j \leqslant |y| - 1$ and $j < l \leqslant |y|$. For each partial ordinal classifier cascade $h_{i,j}$ and for each set of samples $\mathbf{X}$

$$p_h(y_{(l)} \,|\, \mathbf{X}) \leqslant p_{h_{i,j}}(y_{(j+1)} \,|\, \mathbf{X}). \tag{A2}$$

Proof. Theorem A1 states that a partial cascade $h_{i,j}$ predicts its highest class label $y_{(j+1)}$ more often than the full cascade h predicts any subsequent class label $y_{(l)}$, $l \geqslant j + 1$. Let in the following M denote the number of evaluations of the chosen type of resampling experiment on $\mathbf{X}$. The number of samples (more precisely instances) of $\mathbf{X}$ classified as $y_{(l)}$ by the full cascade is then given by

$$m_{(l)} = M \cdot p_h(y_{(l)} \,|\, \mathbf{X}). \tag{A3}$$

The corresponding number for the partial ordinal cascade will be denoted as

$$\dot{m}_{(l)} = M \cdot p_{h_{i,j}}(y_{(l)} \,|\, \mathbf{X}). \tag{A4}$$

Case $i = 1$: If both cascades start with the classifier $c_{(1)}$, the full cascade can be seen as an extension of the partial cascade. The same samples will be classified as $y_{(1)}, \ldots, y_{(j)}$. For these classes, $m_{(1)} = \dot{m}_{(1)}$, …, $m_{(j)} = \dot{m}_{(j)}$. The remaining samples will be classified as $y_{(j+1)}$ by the partial cascade while the full cascade is able to categorise the remaining samples as $y_{(j+1)}, \ldots, y_{(|y|)}$

$$\dot{m}_{(j+1)} = M - \sum_{l=1}^{j} m_{(l)} = \sum_{l=j+1}^{|y|} m_{(l)}. \tag{A5}$$

As all $m_{(i)} \geqslant 0$, $\dot{m}_{(j+1)} = m_{(l)}$, holds for at most one l if the full cascade constantly predicts all $y_{(l)}$ for all remaining samples, which is trivially fulfilled for $j + 1 = |y|$.

For a non constant classification of the samples $\dot{m}_{(j+1)} > m_{(l)}$:
Case $i > 1$: In this case we cannot assume that each sample that is classified as $y_{(i)}, \ldots, y_{(j+1)}$ by the partial cascade receives the same class label by the full cascade. However, if a sample passes classes $y_{(1)}, \ldots, y_{(i-1)}$ in the full cascade they will receive class label $y_{(l)}$ if and only if the partial cascade predicts $y_{(l)}$. None of the base classifiers $c_{(1)}, \ldots, c_{(i-1)}$ will predict a class label $y_{(l)}$, $l \geqslant j+1$. That is, we can give a lower bound on the number of samples that will receive a class label $y_{(1)}, \ldots, y_{(j)}$ from the full cascade by the number of samples that receive a class label $y_{(i)}, \ldots, y_{(j)}$ from the partial cascade

$$\sum_{l=i}^{j} \dot{m}_{(l)} \leqslant \underbrace{\sum_{l=1}^{i-1} m_{(l)}}_{\geqslant 0} + \sum_{l=i}^{j} m_{(l)} \quad \Longleftrightarrow \tag{A6}$$

$$\Longleftrightarrow \quad M - \sum_{l=i}^{j} \dot{m}_{(l)} \geqslant M - \left(\sum_{l=1}^{i-1} m_{(l)} + \sum_{l=i}^{j} m_{(l)} \right) \tag{A7}$$

$$\Longleftrightarrow \quad \dot{m}_{(j+1)} \geqslant \sum_{l=j+1}^{|y|} m_{(l)} \,. \tag{A8}$$

Similar as before, $m_{(i)} \geqslant 0, \dot{m}_{(j+1)} = m_{(l)}$ holds for at most one l, if the classifier omits the prediction of class labels

$$y \in \left\{ y_{(1)}, \ldots, y_{(i-1)}, y_{(j+1)}, \ldots, y_{(|y|)} \right\} \setminus \left\{ y_{(l)} \right\} . \tag{A9}$$

Lemma A1. Let h denote a full ordinal classifier cascade

$$h : \mathbb{R}^n \longrightarrow y = \left\{ y_{(1)}, \ldots, y_{(|y|)} \right\} . \tag{A10}$$

with base classifiers $\mathcal{E} = \left\{ c_{(1)}, \ldots, c_{(|y|-1)} \right\}$. For each $c_{(i)} \in \mathcal{E}$ and for each l with $i < l \leqslant |y|$

$$p_h(y_{(l)} \,|\, \mathbf{X}) \leqslant p_{c_{(i)}}(y_{(i+1)} \,|\, \mathbf{X}). \tag{A11}$$

Proof. Lemma A1 is a direct consequence of Theorem A1 as $c_{(i)} = h_{i,i}$.

Theorem A2. Let h denote a full ordinal classifier cascade

$$h : \mathbb{R}^n \longrightarrow y = \left\{ y_{(1)}, \ldots, y_{(|y|)} \right\} . \tag{A12}$$

Let furthermore i,j,l denote indices with $1 \leqslant i \leqslant j \leqslant |y| - 1$ and $i \leqslant l \leqslant j$. For each partial ordinal classifier cascade $h_{i,j}$ and for each set of samples $\mathbf{X}$

$$p_h(y_{(l)} \,|\, \mathbf{X}) \leqslant p_{h_{i,j}}(y_{(l)} \,|\, \mathbf{X}). \tag{A13}$$

Proof. Theorem A2 states that each class label $y_{(l)}$, $i \leqslant l \leqslant j$ which is covered by a partial cascade $h_{i,j}$ is predicted at least in the same number of cases by the partial cascade as it is by the full cascade. Let $m_{(l)}$ and $\dot{m}_{(l)}$ be defined as in Equations A3 and A4. As before, we can assume that both the full cascade and the partial cascade will classify the same samples as $y_{(i)}, \ldots, y_{(j)}$ if the full cascade did not predict class labels $y_{(1)}, \ldots, y_{(i-1)}$. The set of remaining samples of the partial cascade is therefore a superset of remaining samples of the full cascade

$$M - \sum_{l=i}^{l-1} \dot{m}_{(l)} \geqslant M - \left(\sum_{l=1}^{i-1} m_{(l)} + \sum_{l=i}^{l-1} m_{(l)} \right). \tag{A14}$$

The additional remaining samples of the partial ordinal cascade can increase the number of predictions of class $y_{(l)}$.

Lemma A2. Let h denote an full ordinal classifier cascade

$$h : \mathbb{R}^n \longrightarrow \mathcal{Y} = \left\{ y_{(1)}, \ldots, y_{(|\mathcal{Y}|)} \right\}. \tag{A15}$$

with base classifiers $\mathcal{E} = \left\{ c_{(1)}, \ldots, c_{(|\mathcal{Y}|-1)} \right\}$. For each $c_{(i)} \in \mathcal{E}$ and for each $l \in \{i, i+1\}$

$$p_h(y_{(l)} \,|\, \mathbf{X}) \leqslant p_{c_{(i)}}(y_{(l)} \,|\, \mathbf{X}). \tag{A16}$$

Proof. Lemma A2 is a direct consequence of Theorem A2 as $c_{(i)} = h_{i,i}$.

Performance relations between cascades and their subcascades (Section 3.1.3)

Proof. Subcascade-cascade performance relations [122]
There is a monotonicity on the minimal class-wise sensitivity on the number of classes

$$\text{smin}_h(\mathcal{Y}) \leqslant \text{smin}_h(\mathcal{Y}'), \tag{A17}$$

for $\mathcal{Y}' \subset \mathcal{Y}, \mathcal{Y}' \neq \emptyset$.

This proof was published and is taken from Lausser* et al. [122].
Cascade h' is constructed from the same base classifiers as h. Only one classifier is removed, which closes the option of predicting class label $y_{(i)}$. In case $i = 1$, all samples that would have been classified as $y_{(1)}$ by h are directly sent to $c_{(2)}$ by h'. This is of course also true for all remaining samples with $c_{(1)}(\mathbf{x}) = y_{(2)}$ in h. Therefore for each class label $y \in \mathcal{Y}'$, the set of samples that receive this class

label y by h' comprises at least all samples that receive class label y by h. The corresponding class-wise sensitivities and their minimum can only be increased.
In case $1 < i < |\mathcal{Y}|$, additionally classifier $c_{(i-1)}$ and $c_{(i)}$ are replaced by $c'_{(i-1)}$, which redirects those samples that would be sent from $c_{(i-1)}$ to $c_{(i)}$. The common preamble of base classifiers of h and h' leads to identical class-wise sensitivities for classes $y_{(1)}, \ldots, y_{(i-1)}$. Those classes that received class label $y_{(i)}$ by h are again reclassified into one of the subsequent classes leading to equivalent or increased class-wise sensitivities.
In case $i = |\mathcal{Y}|$, the cascades h and h' share the longest possible preamble of base classifiers. Only the samples receiving class label $y_{(|\mathcal{Y}|)}$ by h will be reclassified as $y_{(|\mathcal{Y}|-1)}$ by h' leading to at least the class-wise sensitivity for class $y_{(|\mathcal{Y}|)}$.
We therefore get for all cases

$$\operatorname{smin}_h(\mathcal{Y}') \leqslant \operatorname{smin}_{h'}(\mathcal{Y}'). \tag{A18}$$

Ordinal classifier cascade - dependency on dimensionality (Section 3.2)

Proof. Sensitivity bounds on *one-against-one* (OaO) class-wise sensitivity
The decomposition of the conditional prediction rate $p_{h_{\mathrm{OaO}}}(y|y)$ (Equation 37) consists of three parts. Two parts refer to class-wise sensitivities and one to external rates.
As the external rates $p_{\mathrm{rest}}(y, \mathcal{Y}_1, \mathcal{Y}_0) \leqslant 1$, bounds depending only on the binary class-wise sensitivities (cl.sens.) can be given. The upper bound is

$$\begin{aligned} p_{h_{\mathrm{OaO}}}(y|y) &\leqslant \sum_{i=\lceil \frac{|\mathcal{O}|}{2} \rceil}^{|\mathcal{O}|} \sum_{\substack{\mathcal{Y}_1 \subseteq \mathcal{O} \\ |\mathcal{Y}_1|=i}} \left(\prod_{y' \in \mathcal{Y}_1} \underbrace{p_{c_{yy'}}(y|y)}_{\text{cl.sens.}} \right) \left(\prod_{\substack{\mathcal{Y}_0 = \mathcal{O} \setminus \mathcal{Y}_1 \\ y'' \in \mathcal{Y}_0}} 1 - \underbrace{p_{c_{yy''}}(y|y)}_{\text{cl.sens.}} \right) \\ &\leqslant \sum_{i=\lceil \frac{|\mathcal{O}|}{2} \rceil}^{|\mathcal{O}|} \sum_{\substack{\mathcal{Y}_1 \subseteq \mathcal{O} \\ |\mathcal{Y}_1|=i}} \max_{y' \in \mathcal{Y}_1} p_{c_{yy'}}(y|y)^{|\mathcal{Y}_1|} \left(1 - \min_{y'' \in \mathcal{Y}_0} p_{c_{yy''}}(y|y) \right)^{|\mathcal{Y}_0|} \\ &\leqslant \sum_{i=\lceil \frac{|\mathcal{O}|}{2} \rceil}^{|\mathcal{O}|} \binom{|\mathcal{O}|}{i} p_{\max}^{i} (1 - p_{\min})^{|\mathcal{O}|-i}, \end{aligned} \tag{A19}$$

where $p_{\max}$ is the highest and $p_{\min}$ the lowest sensitivity of all base classifiers that have been trained on the class under consideration.

A lower bound on $p_{h_{OaO}}(y|y)$ can be given assuming that all base classifiers that can predict y, correctly predict y ($\mathcal{Y}_1 = \mathcal{O}$). With p_{min} being the smallest sensitivity of all base classifiers that can predict the class under consideration, it holds:

$$p_{h_{OaO}}(y|y) \geqslant \prod_{y' \in \mathcal{O}} \underbrace{p_{c_{yy'}}(y|y)}_{\text{cl.sens.}} \geqslant \min_{y' \in \mathcal{O}} (\underbrace{p_{c_{yy'}}(y|y)}_{\text{cl.sens.}})^{|\mathcal{O}|} \geqslant p_{min}^{|\mathcal{O}|}. \tag{A20}$$

Characterisation of the artificial data (Section 3.3.2)

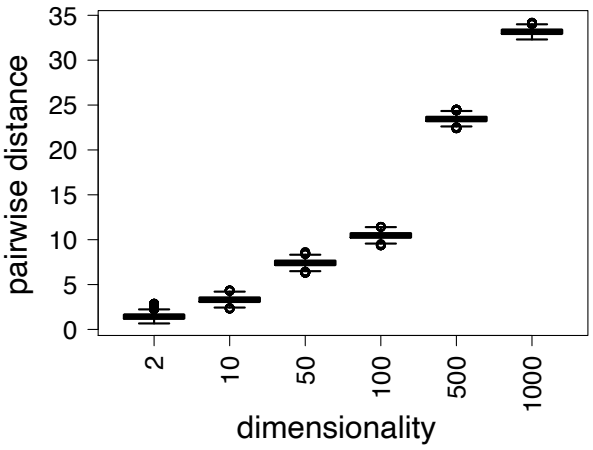


Figure A1: The step size in the prototype sampling procedure depends on the dimensionality. The distances between all direct neighboured prototypes in the patterns with an underlying linear relation is shown. This refers to all distances reflecting one step in the sampling procedure. Each pattern was sampled 100 times. The higher the dimension the larger the step size.

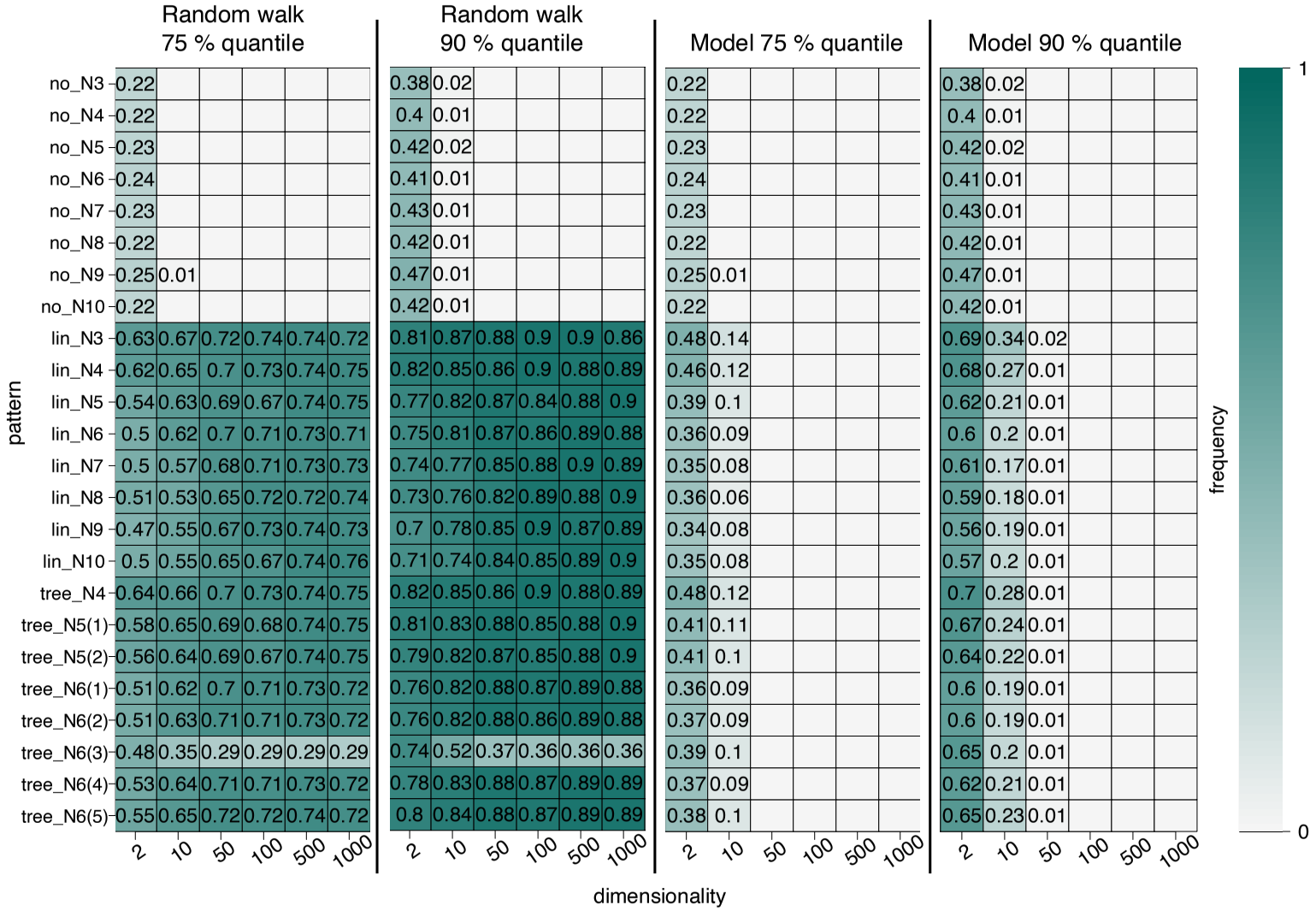


Figure A2: Probability to sample a circle on the sample level. The border characterising the sampling of a circle is given once by the 75 % quantile and by the 90 % quantile of the distances of samples of all classes towards their own prototypes.

Figure A3: Standard deviation of the frequency of not expected cascades. A global threshold of 0.7 was imposed on the screening results. Three different sets of remaining cascades were compared to the set of expected cascades, that is given by the underlying corresponding pattern (upper part). A: All top ranked cascades of all lengths are used. B: Only the longest once passing the threshold and C: the Pareto optimal set is given. Additionally, the standard deviation of the frequency of one often occurring error is shown (lower part), which is a swap in the last two classes of a cascade. The underlying data has a dimensionality of $n = 1000$ and there were 100 datasets sampled per pattern.

Phenotype network reconstruction - Reconstruction results (Section 3.4.4)

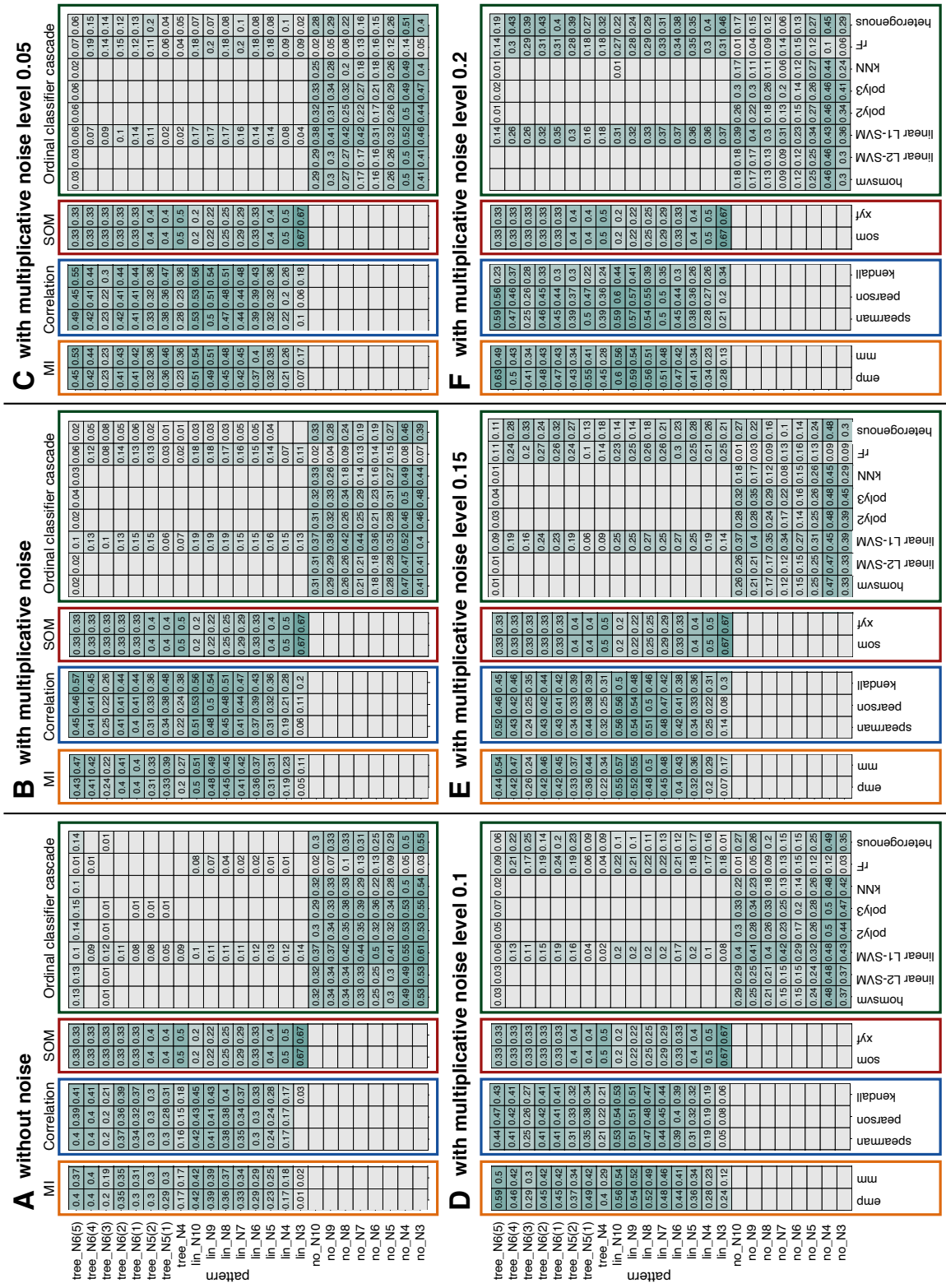


Figure A4: Error rates using a global threshold of 0.3 and varying noise levels. (Continued on the following page.)

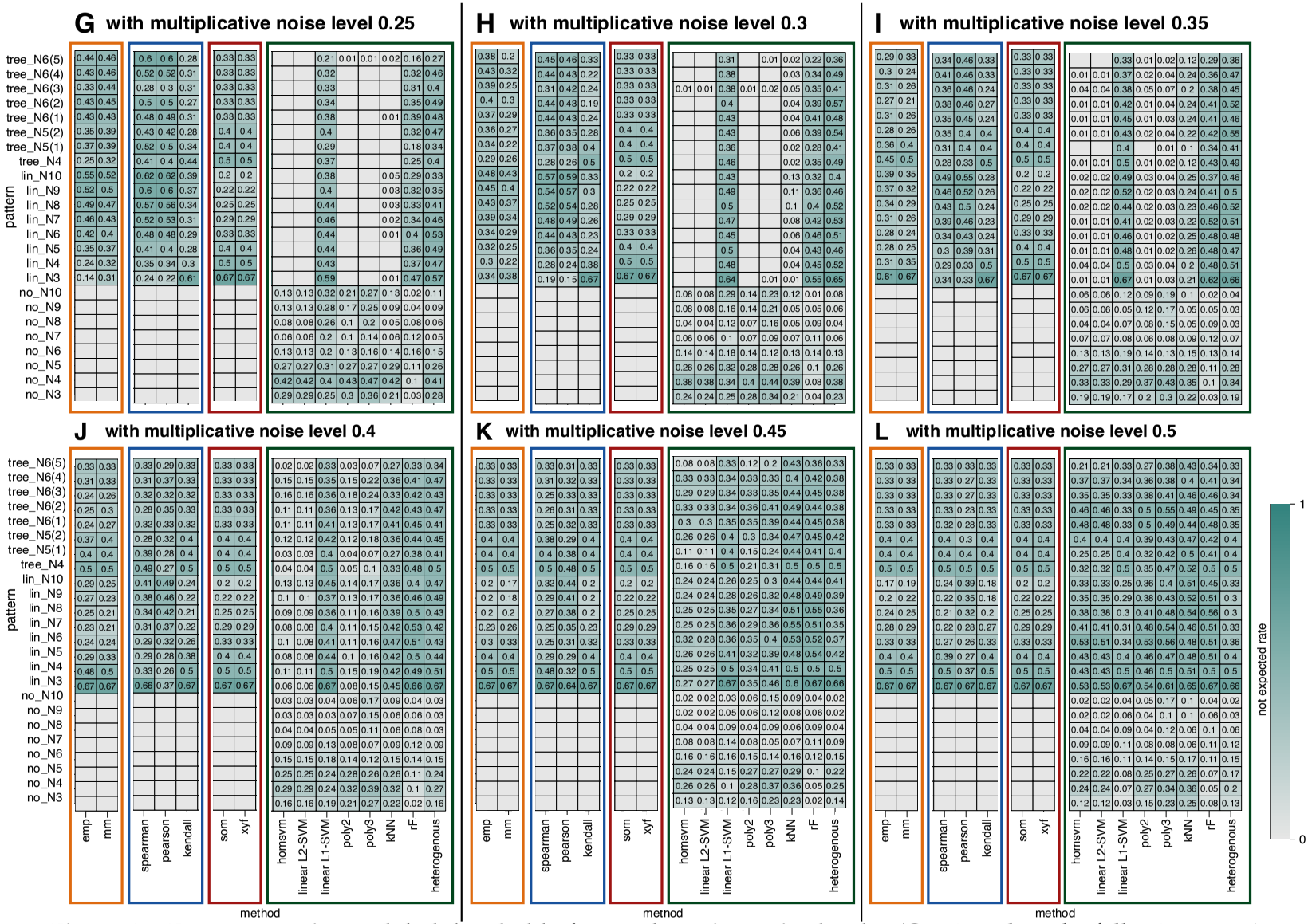


Figure A4: Error rates using a global threshold of 0.3 and varying noise levels. (Continued on the following page.)

Figure A4: (Previous page.) Error rates using a global threshold of 0.3 and varying noise levels. All cascades passing a threshold of 1 are used for reconstruction. The methods are grouped: MI=Mutual information (emp=empirical and mm=Miller-Madow corrected), Correlation (spearman, pearson, kendall), SOM=self-organising map (som= unsupervised version, xyf=supervised version) and ordinal classifier cascade (Occ) (homsvm=linear L2-support vector machine (SVM) with constant cost parameter for all base classifiers, linear L2-SVM, linear L1-SVM, SVM with polynomial degree of 2 and 3, as poly2 and poly3, k-nearest neighbour (kNN), random forest (rF) with parameters optimised in a hyperparameter tuning and heterogenous, which is an ensemble selecting the best binary classifier per step from the set of parameter tuned classifiers). The underlying data has a dimensionality of $n = 1000$ and there were 100 datasets sampled per pattern.

Cascades in developmental stages of model organisms (Section 5.2)

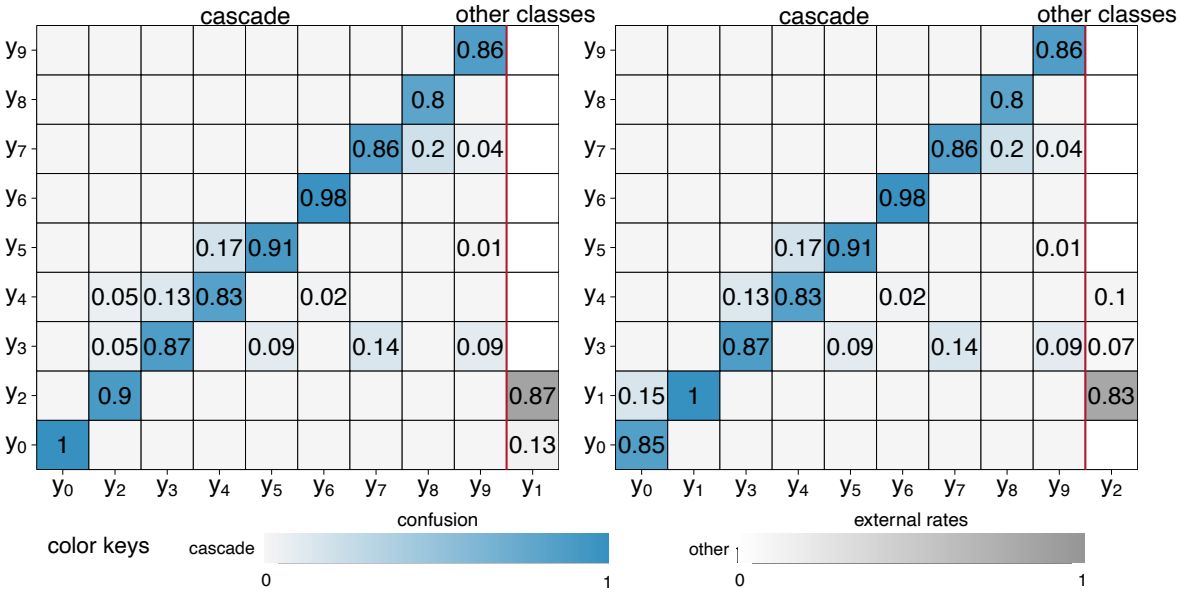
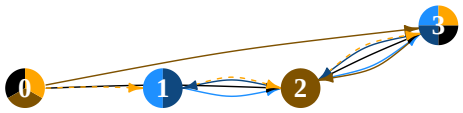
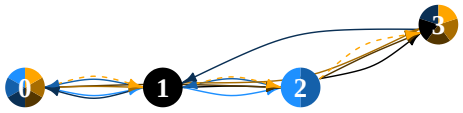
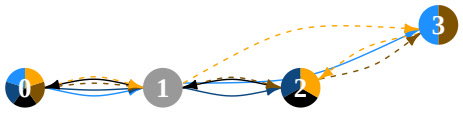
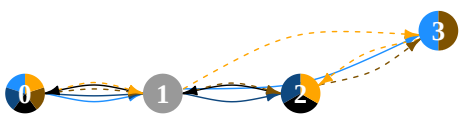
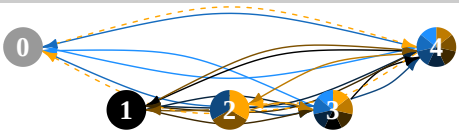
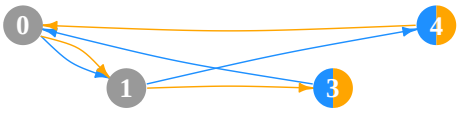
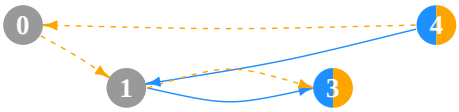
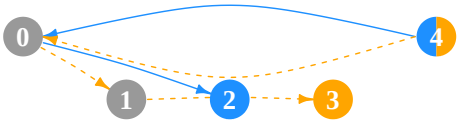


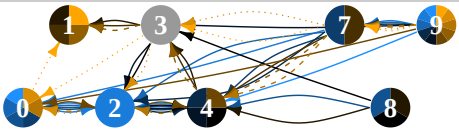
Figure A5: Confusion matrices of the set of returned subcascades of d_6 *Caenorhabditis elegans* (*C. elegans*) time-points. The cascade classes are given in the rows and ordered from bottom to top. The columns comprise all classes and the values correspond to the confusion values and external rates indicating the amount of samples of the class given in the column that is assigned to the class given in the row.

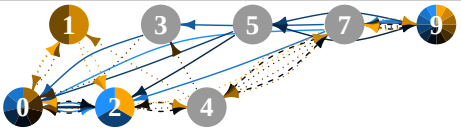
Cascades in datasets with (pre-)cancer stages - Screening results (Section 5.3.3)

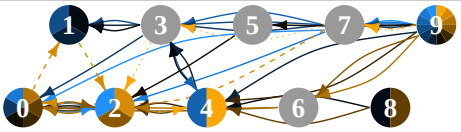
Table A1: Screening results on the datasets selecting the longest cascades. The results for each hyperparameter tuning (binary accuracy, binary sensitivity, cost parameter (C) homogenous) and each base classifier (linear L2-support vector machine (SVM), linear L1-support vector machine (SVM)) is given. As the linear L2-SVM resulted in the same set independent of the optimisation method it is given only once. For each set the performance is denoted by the minimal class-wise sensitivity (smin) in %. Shown are the best longest cascades (blc) and cascades in a sensitivity range of smin $\geqslant \text{smin}_{\text{blc}} - 0.05$ around the performance of the best longest cascade smin_{blc}. The graphical illustration shows an arrangement of all cascades found. Each cascade is represented by one colour. The different lengths are given by the line type. Nodes that indicate the first position within a cascade are coloured based on the corresponding cascade, of that this node represents the first position. Nodes that are not used at the first position are coloured grey and cascades that are included independent of the base classifier and tuning strategy are marked boldface.

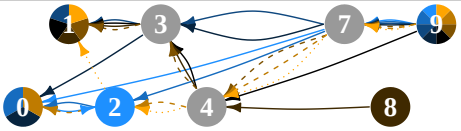
d_7: Colon carcinoma			
cascades	smin [%]		graph
L2-SVM, all methods			
$y_1 \prec y_2 \prec y_3$, $y_3 \prec y_2 \prec y_1$, $y_0 \prec y_2 \prec y_3$, $y_0 \prec y_3 \prec y_2$	$= 100$	—	
$y_0 \prec y_1 \prec y_2 \prec y_3$	$= 94.1$	- - -	
L1-SVM, accuracy tuned			
$y_0 \prec y_1 \prec y_2$, $y_2 \prec y_1 \prec y_0$, $y_3 \prec y_1 \prec y_0$, $y_1 \prec y_2 \prec y_3$, $y_0 \prec y_2 \prec y_3$, $y_0 \prec y_1 \prec y_3$	$\geqslant 90.9$	—	
$y_0 \prec y_1 \prec y_2 \prec y_3$	$= 81.8$	- - -	
L1-SVM, sensitivity tuned			
$y_0 \prec y_1 \prec y_3$, $y_0 \prec y_1 \prec y_2$, $y_2 \prec y_1 \prec y_0$	$\geqslant 94.1$	—	
$y_0 \prec y_1 \prec y_2 \prec y_3$, $y_0 \prec y_1 \prec y_3 \prec y_2$	$= 72.7$	- - -	
L1-SVM, homogenous			
$y_0 \prec y_1 \prec y_3$, $y_0 \prec y_1 \prec y_2$, $y_2 \prec y_1 \prec y_0$	$\geqslant 94.1$	—	
$y_0 \prec y_1 \prec y_2 \prec y_3$, $y_0 \prec y_1 \prec y_3 \prec y_2$	$= 72.7$	- - -	

d_8: Pancreatic adenocarcinoma (PDAC), Microarray			
cascades	smin [%]		graph
L2-SVM, all methods			
$y_3 \prec y_0 \prec y_4$, $y_4 \prec y_0 \prec y_3$, $y_2 \prec y_3 \prec y_4$, $y_3 \prec y_1 \prec y_4$, $y_3 \prec y_4 \prec y_1$, $y_4 \prec y_1 \prec y_3$, $y_2 \prec y_1 \prec y_4$, $y_4 \prec y_2 \prec y_3$	$\geqslant 80$	—	
$y_3 \prec y_0 \prec y_4 \prec y_2$	$= 75$	- - -	
L1-SVM, accuracy tuned			
$y_3 \prec y_0 \prec y_1 \prec y_4$, $y_4 \prec y_0 \prec y_1 \prec y_3$	$= 83.3$	—	
L1-SVM, sensitivity tuned			
$y_4 \prec y_1 \prec y_3$	$\geqslant 90$	—	
$y_4 \prec y_0 \prec y_1 \prec y_3$	$= 83.3$	- - -	
L1-SVM, homogenous			
$y_4 \prec y_0 \prec y_2$	$= 100$	—	
$y_4 \prec y_0 \prec y_1 \prec y_3$	$= 83.3$	- - -	

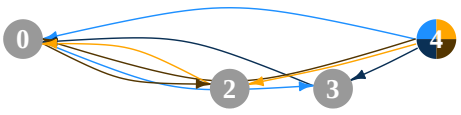
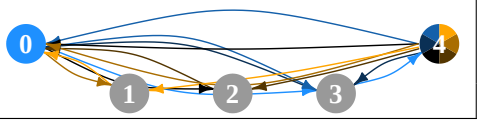
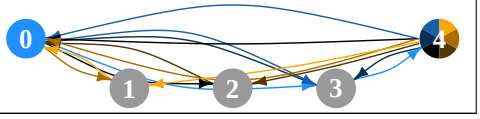
d_9: Hepatocellular carcinoma (HCC)			
cascades	smin [%]		graph
L2-SVM, all methods			
$y_9 \prec y_4 \prec y_2 \prec y_0$, $\mathbf{y_9 \prec y_7 \prec y_0 \prec y_2}$, $\mathbf{y_9 \prec y_7 \prec y_2 y_0}$, $y_8 \prec y_4 \prec y_2 \prec y_0$, $y_7 \prec y_4 \prec y_2 \prec y_0$, $y_9 \prec y_7 \prec y_2 \prec y_4$, $y_9 \prec y_3 \prec y_2 \prec y_4$, $y_8 \prec y_3 \prec y_2 \prec y_4$, $y_8 \prec y_4 \prec y_3 \prec y_1$, $y_7 \prec y_4 \prec y_3 \prec y_1$, $y_9 \prec y_0 \prec y_2 \prec y_4$	$\geqslant 83.3$	—	
$y_9 \prec y_7 \prec y_4 \prec y_3 \prec y_1$, $y_9 \prec y_7 \prec y_4 \prec y_2 \prec y_0$	$= 77.8$	- - -	
$y_9 \prec y_7 \prec y_3 \prec y_2 \prec y_0$, $y_9 \prec y_7 \prec y_3 \prec y_0 \prec y_1$	$\geqslant 75$	·····	

L1-SVM, accuracy tuned			
$y_9 \prec y_7 \prec y_0 \prec y_2$, $y_9 \prec y_7 \prec y_2 \prec y_0$, $y_9 \prec y_7 \prec y_3 \prec y_0$, $y_9 \prec y_5 \prec y_0 \prec y_2$, $y_9 \prec y_5 \prec y_2 \prec y_0$	$\geqslant 83.3$	—	
$y_0 \prec y_2 \prec y_4 \prec y_7 \prec y_9$, $y_9 \prec y_7 \prec y_4 \prec y_2 \prec y_0$	$= 77.8$	- - -	
$y_9 \prec y_7 \prec y_4 \prec y_3 \prec y_0$, $y_9 \prec y_7 \prec y_4 \prec y_0 \prec y_1$, $y_9 \prec y_7 \prec y_4 \prec y_1 \prec y_0$, $y_1 \prec y_2 \prec y_4 \prec y_7 \prec y_9$, $y_9 \prec y_7 \prec y_0 \prec y_1 \prec y_2$	$\geqslant 75$	·····	

L1-SVM, sensitivity tuned			
$y_9 \prec y_7 \prec y_0 \prec y_2$, $y_9 \prec y_7 \prec y_2 \prec y_0$, $y_9 \prec y_7 \prec y_3 \prec y_4$, $y_9 \prec y_7 \prec y_3 \prec y_1$, $y_9 \prec y_7 \prec y_3 \prec y_0$, $y_9 \prec y_4 \prec y_3 \prec y_1$, $y_8 \prec y_4 \prec y_3 \prec y_1$, $y_9 \prec y_5 \prec y_2 \prec y_0$, $y_9 \prec y_4 \prec y_2 \prec y_0$, $y_8 \prec y_4 \prec y_2 \prec y_0$, $y_9 \prec y_6 \prec y_0 \prec y_2$, $y_9 \prec y_6 \prec y_2 \prec y_0$	$\geqslant 80$	—	
$y_9 \prec y_7 \prec y_0 \prec y_1 \prec y_2$	$= 75$	- - -	
$y_9 \prec y_7 \prec y_3 \prec y_2 \prec y_4$	$\geqslant 72.7$	·····	

L1-SVM, homogenous			
$y_9 \prec y_7 \prec y_0 \prec y_2$, $y_9 \prec y_7 \prec y_2 \prec y_0$, $y_9 \prec y_7 \prec y_3 \prec y_1$, $y_9 \prec y_7 \prec y_3 \prec y_0$, $y_9 \prec y_4 \prec y_3 \prec y_1$, $y_8 \prec y_4 \prec y_3 \prec y_1$	$\geqslant 83.3$	—	
$y_9 \prec y_7 \prec y_4 \prec y_3 \prec y_1$, $y_9 \prec y_7 \prec y_4 \prec y_2 \prec y_0$	$= 77.8$	- - -	
$y_9 \prec y_7 \prec y_4 \prec y_2 \prec y_1$	$\geqslant 75$	·····	

d_{10}: PDAC, RNA-Sequencing

cascades	smin [%]		graph
L2-SVM, all methods			
$y_4 \prec y_0 \prec y_3$, $y_4 \prec y_3 \prec y_0$, $y_4 \prec y_0 \prec y_2$, $y_4 \prec y_2 \prec y_0$	$= 100$	—	
L1-SVM, accuracy tuned			
$y_0 \prec y_3 \prec y_4$, $y_4 \prec y_0 \prec y_3$, $y_4 \prec y_3 \prec y_0$, $y_4 \prec y_0 \prec y_2$, $y_4 \prec y_2 \prec y_0$, $y_4 \prec y_0 \prec y_1$, $y_4 \prec y_1 \prec y_0$	$= 100$	—	
L1-SVM, sensitivity tuned			
$y_0 \prec y_3 \prec y_4$, $y_4 \prec y_0 \prec y_3$, $y_4 \prec y_3 \prec y_0$, $y_4 \prec y_0 \prec y_2$, $y_4 \prec y_2 \prec y_0$, $y_4 \prec y_0 \prec y_1$, $y_4 \prec y_1 \prec y_0$	$= 100$	—	

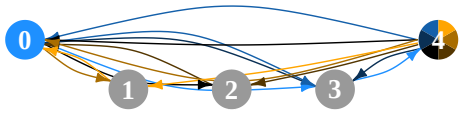
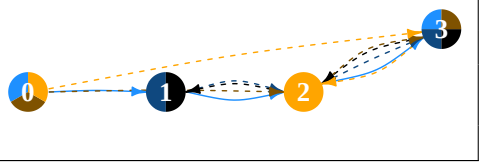
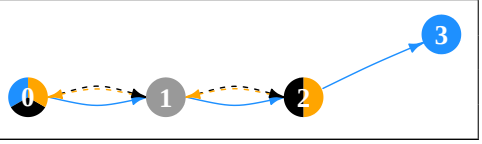
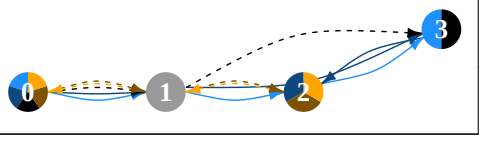
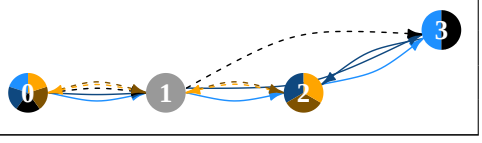
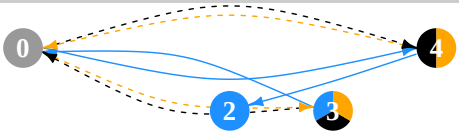
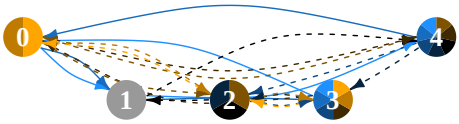
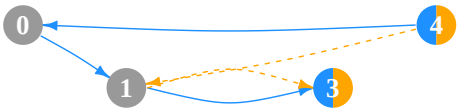
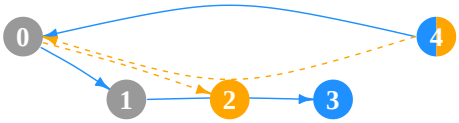
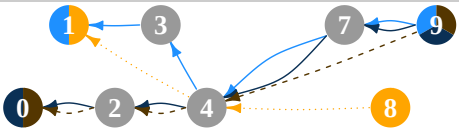
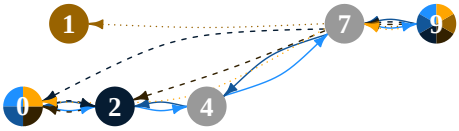
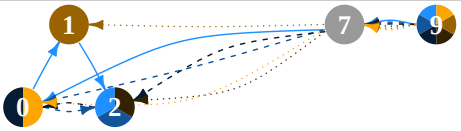
L1-SVM, homogenous			
$y_0 \prec y_3 \prec y_4$, $y_4 \prec y_0 \prec y_3$, $y_4 \prec y_3 \prec y_0$, $y_4 \prec y_0 \prec y_2$, $y_4 \prec y_2 \prec y_0$, $y_4 \prec y_0 \prec y_1$, $y_4 \prec y_1 \prec y_0$	$= 100$	—	

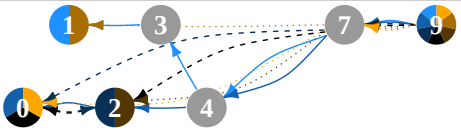
Table A2: Screening results on the datasets selecting the Pareto optimal solutions. The results and the minimal class-wise sensitivity (smin [%]) for each hyperparameter tuning (binary accuracy, binary sensitivity, cost parameter (C) homogenous) and each base classifier (linear L2-support vector machine (SVM), linear L1-SVM) is given. As the linear L2-SVM resulted in the same set independent of the optimisation method it is given only once. The graphical illustration shows an arrangement of all cascades found. Each cascade is represented by one colour. The different lengths are given by the line type. Nodes that indicate the first position within a cascade are coloured based on the corresponding cascade, of that this node represents the first position. Nodes that are not used at the first position are coloured grey and cascades that are included independent of the base classifier and tuning strategy are marked boldface.

d_7: Colon carcinoma			
cascades	smin [%]		graph
L2-SVM, all methods			
$\mathbf{y_0 \prec y_1 \prec y_2 \prec y_3}$	$= 94.1$	—	
$y_1 \prec y_2 \prec y_3$, $y_3 \prec y_2 \prec y_1$, $y_0 \prec y_2 \prec y_3$, $y_0 \prec y_3 \prec y_2$	$= 100$	- - -	
L1-SVM, accuracy tuned			
$\mathbf{y_0 \prec y_1 \prec y_2 \prec y_3}$	$= 81.8$	—	
$y_0 \prec y_1 \prec y_2$, $y_2 \prec y_1 \prec y_0$	$= 94.1$	- - -	
L1-SVM, sensitivity tuned			
$\mathbf{y_0 \prec y_1 \prec y_2 \prec y_3}$, $y_0 \prec y_1 \prec y_3 \prec y_2$	$= 72.7$	—	
$y_0 \prec y_1 \prec y_3$, $y_0 \prec y_1 \prec y_2$, $y_2 \prec y_1 \prec y_0$	$= 94.1$	- - -	
L1-SVM, homogenous			
$\mathbf{y_0 \prec y_1 \prec y_2 \prec y_3}$, $y_0 \prec y_1 \prec y_3 \prec y_2$	$= 72.7$	—	
$y_0 \prec y_1 \prec y_3$, $y_0 \prec y_1 \prec y_2$, $y_2 \prec y_1 \prec y_0$	$= 94.1$	- - -	

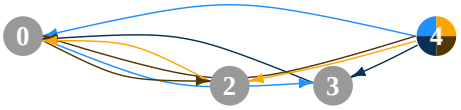
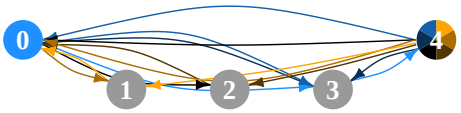
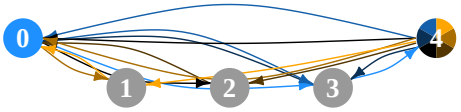
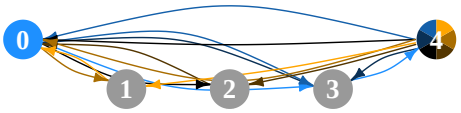
d_8: Pancreatic adenocarcinoma (PDAC), Microarray			
cascades	smin [%]		graph
L2-SVM, all methods			
$y_3 \prec y_0 \prec y_4 \prec y_2$	$= 75$	—	
$y_3 \prec y_0 \prec y_4$, $y_4 \prec y_0 \prec y_3$	$= 90$	- - -	
L1-SVM, accuracy tuned			
$y_3 \prec y_0 \prec y_1 \prec y_4$, $y_4 \prec y_0 \prec y_1 \prec y_3$	$= 83.3$	—	
$y_4 \prec y_2 \prec y_3$, $y_4 \prec y_3 \prec y_2$, $y_2 \prec y_1 \prec y_4$, $y_3 \prec y_0 \prec y_4$, $y_4 \prec y_0 \prec y_2$, $y_0 \prec y_2 \prec y_3$, $y_3 \prec y_2 \prec y_0$	$= 87.5$	- - -	
L1-SVM, sensitivity tuned			
$y_4 \prec y_0 \prec y_1 \prec y_3$	$= 83.3$	—	
$y_4 \prec y_1 \prec y_3$	$= 90$	- - -	
L1-SVM, homogenous			
$y_4 \prec y_0 \prec y_1 \prec y_3$	$= 83.3$	—	
$y_4 \prec y_0 \prec y_2$	$= 100$	- - -	

d_9: Hepatocellular carcinoma (HCC)			
cascades	smin [%]		graph
L2-SVM, all methods			
$y_9 \prec y_7 \prec y_4 \prec y_3 \prec y_1$, $y_9 \prec y_7 \prec y_4 \prec y_2 \prec y_0$	$= 77.8$	—	
$y_9 \prec y_4 \prec y_2 \prec y_0$	$= 90.9$	- - -	
$y_8 \prec y_4 \prec y_1$	$= 100$	·····	
L1-SVM, accuracy tuned			
$y_0 \prec y_2 \prec y_4 \prec y_7 \prec y_9$, $y_9 \prec y_7 \prec y_4 \prec y_2 \prec y_0$	$= 77.8$	—	
$y_9 \prec y_7 \prec y_0 \prec y_2$, $y_9 \prec y_7 \prec y_2 \prec y_0$	$= 84.6$	- - -	
$y_9 \prec y_7 \prec y_1$, $y_9 \prec y_7 \prec y_0$	$= 100$	·····	

L1-SVM, sensitivity tuned			
$y_9 \prec y_7 \prec y_0 \prec y_1 \prec y_2$	$= 75$	—	
$y_9 \prec y_7 \prec y_0 \prec y_2$, $y_9 \prec y_7 \prec y_2 \prec y_0$	$= 92.3$	- - -	
$y_9 \prec y_7 \prec y_2$, $y_9 \prec y_7 \prec y_1$, $y_9 \prec y_7 \prec y_0$	$= 100$	·····	

L1-SVM, homogenous			
$y_9 \prec y_7 \prec y_4 \prec y_3 \prec y_1$, $y_9 \prec y_7 \prec y_4 \prec y_2 \prec y_0$	$= 77.8$	—	
$y_9 \prec y_7 \prec y_0 \prec y_2$, $y_9 \prec y_7 \prec y_2 \prec y_0$	$= 92.3$	- - -	
$y_9 \prec y_7 \prec y_2$, $y_9 \prec y_7 \prec y_1$, $y_9 \prec y_7 \prec y_0$	$= 100$	·····	

d_{10}: PDAC, RNA-Sequencing

cascades	smin [%]		graph
L2-SVM, all methods			
$y_4 \prec y_0 \prec y_3$, $y_4 \prec y_3 \prec y_0$, $y_4 \prec y_0 \prec y_2$, $y_4 \prec y_2 \prec y_0$	$= 100$	—	
L1-SVM, accuracy tuned			
$y_0 \prec y_3 \prec y_4$, $y_4 \prec y_0 \prec y_3$, $y_4 \prec y_3 \prec y_0$, $y_4 \prec y_0 \prec y_2$, $y_4 \prec y_2 \prec y_0$, $y_4 \prec y_0 \prec y_1$, $y_4 \prec y_1 \prec y_0$	$= 100$	—	
L1-SVM, sensitivity tuned			
$y_0 \prec y_3 \prec y_4$, $y_4 \prec y_0 \prec y_3$, $y_4 \prec y_3 \prec y_0$, $y_4 \prec y_0 \prec y_2$, $y_4 \prec y_2 \prec y_0$, $y_4 \prec y_0 \prec y_1$, $y_4 \prec y_1 \prec y_0$	$= 100$	—	
L1-SVM, homogenous			
$y_0 \prec y_3 \prec y_4$, $y_4 \prec y_0 \prec y_3$, $y_4 \prec y_3 \prec y_0$, $y_4 \prec y_0 \prec y_2$, $y_4 \prec y_2 \prec y_0$, $y_4 \prec y_0 \prec y_1$, $y_4 \prec y_1 \prec y_0$	$= 100$	—	

Confirming carcinoma type relatedness (Section 5.3.4.3)

Table A3: Results on Kyoto Encyclopedia of Genes and Genomes (KEGG) gene enrichment analysis of d_9 hepatocellular carcinoma dataset. The top 15 results with the lowest adjusted p-value for each of the two finally selected cascades are reported and compared to results from other studies. The characteristics given are: the number of proteins within each pathway N°, the number of proteins included in the set (hits), the false discovery rate (FDR) adjusted p-value, the terms' id and description and references to publications that have found the same pathway by performing a gene enrichment analysis on their sets of differential expressed genes.

term id	N° proteins	hits	adjusted p-value	term description	reference
d_9: Hepatocellular carcinoma $y_9 \prec y_7 \prec y_4 \prec y_3 \prec y_1$					
1100	1161	165	1.0E-53	Metabolic pathways	[222]
5204	71	28	2.3E-20	Chemical carcinogenesis	[72, 86, 128, 222, 140]
982	63	24	4.3E-17	Drug metabolism - cytochrome P450	[234, 86, 237, 128, 222, 247, 140]
4151	336	49	2.1E-15	PI3K-Akt signaling pathway	[234, 45]
980	67	23	2.1E-15	Metabolism of xenobiotics by cytochrome P450	[86, 237, 128, 247, 191, 140]
4060	260	41	4.0E-14	Cytokine-cytokine receptor interaction	[234, 102]
830	58	20	1.6E-13	Retinol metabolism	[234, 72, 86, 237, 128, 222, 247, 191, 140]
5200	320	44	4.4E-13	Pathways in cancer	[244]
4110	121	26	3.0E-12	Cell cycle	[234, 86, 244, 237, 128, 222, 140]
4974	84	22	3.0E-12	Protein digestion and absorption	[199, 191]
5202	164	30	3.1E-12	Transcriptional misregulation in cancer	-
4510	202	33	4.8E-12	Focal adhesion	[248, 199, 140]
4068	120	25	1.4E-11	FoxO signaling pathway	[45]
4610	68	19	2.9E-11	Complement and coagulation cascades	[234, 72, 248, 237, 128, 222, 191]
1230	70	19	4.8E-11	Biosynthesis of amino acids	[128]
d_9: Hepatocellular carcinoma $y_9 \prec y_7 \prec y_4 \prec y_2 \prec y_0$					
1100	1161	223	7.7E-79	Metabolic pathways	[222]
5204	71	33	1.7E-23	Chemical carcinogenesis	[72, 86, 128, 222, 140]
982	63	31	3.8E-23	Drug metabolism - cytochrome P450	[234, 86, 237, 128, 222, 247, 140]
980	67	31	3.1E-22	Metabolism of xenobiotics by cytochrome P450	[86, 237, 128, 247, 191, 140]
830	58	25	5.4E-17	Retinol metabolism	[234, 72, 86, 237, 128, 222, 247, 191, 140]
5202	164	39	4.3E-16	Transcriptional misregulation in cancer	-
5200	320	55	5.9E-16	Pathways in cancer	[244]
5146	104	31	7.2E-16	Amoebiasis	[140]
4974	84	28	8.4E-16	Protein digestion and absorption	[199, 191]
4151	336	55	3.8E-15	PI3K-Akt signaling pathway	[234, 45]
4610	68	23	3.4E-13	Complement and coagulation cascades	[234, 72, 248, 237, 128, 222, 191]
1230	70	23	6.3E-13	Biosynthesis of amino acids	[128]
4060	260	44	1.0E-12	Cytokine-cytokine receptor interaction	[234, 102]
4066	103	27	1.6E-12	HIF-1 signaling pathway	-
40	33	16	3.4E-12	Pentose and glucuronate interconversions	[128]

PDAC RNA-Sequencing (d_{10}) analysis results (Section 5.3)

Within this section the results of dataset d_{10} (pancreatic adenocarcinoma (PDAC) RNA-Sequencing) are given. The base classifier performance of homogenous ordinal classifier cascade (hOcc) with the optimised hyperparameter is shown in Figure A6. The robustness of the L1-SVM hOcc against the cost parameter is given in Table A4. Furthermore, Figure A7 shows the extended confusion matrices of all longest cascades of the Pareto optimal set using L1-SVM hOcc. Figure A8 denotes characteristics of the selected features for all analysed cascades, once the frequency distribution of these features (A) and also the overlap of the features selected by the individual base classifiers (B).

Table A4: Robustness of returned cascades against the cost parameter using L1-support vector machine (SVM) homogenous ordinal classifier cascade (hOcc) on the pancreatic adenocarcinoma (PDAC) RNA-Sequencing dataset (d_{10}). The cascades that were selected for more than 10 of the 37 (0.1,0.2, ... 1, 2, ..., 10, 15, ..., 100) different cost parameters are reported for two different selection methods. Once the Pareto optimal solutions are selected and once all cascades passing a minimal class-wise sensitivity of 0.7. The cascades that were included in the final set for the chosen cost parameter ($C = 0.1$) are marked in boldface.

Pareto optimal set		Set of longest cascades $t \geqslant 0.7$	
PDAC RNA-Sequencing	N°	PDAC RNA-Sequencing	N°
$\mathbf{y_4 \prec y_0 \prec y_1}$	37	$\mathbf{y_4 \prec y_2 \prec y_0}$	37
$\mathbf{y_4 \prec y_2 \prec y_0}$	36	$\mathbf{y_4 \prec y_1 \prec y_0}$	37
$\mathbf{y_4 \prec y_0 \prec y_2}$	36	$\mathbf{y_4 \prec y_0 \prec y_2}$	37
$\mathbf{y_4 \prec y_1 \prec y_0}$	35	$\mathbf{y_4 \prec y_0 \prec y_1}$	37
$\mathbf{y_4 \prec y_0 \prec y_3}$	23	$\mathbf{y_4 \prec y_0 \prec y_3}$	23
$\mathbf{y_4 \prec y_3 \prec y_0}$	22	$\mathbf{y_0 \prec y_3 \prec y_4}$	23
$\mathbf{y_0 \prec y_3 \prec y_4}$	15	$\mathbf{y_4 \prec y_3 \prec y_0}$	22
		$y_0 \prec y_1 \prec y_4$	19

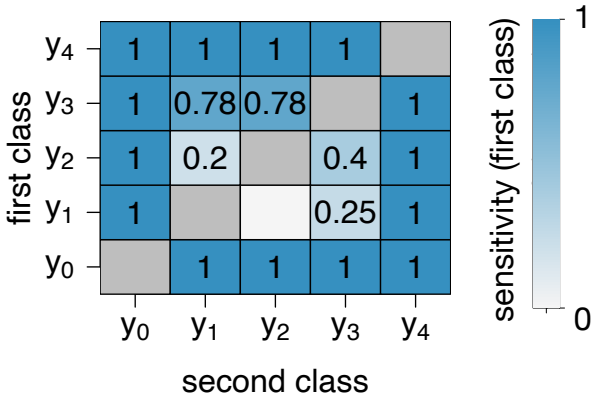


Figure A6: Base classifier performance of L1-support vector machine (SVM) homogenous ordinal classifier cascade (hOcc) on the pancreatic adenocarcinoma (PDAC) RNA-Sequencing dataset (d_{10}). The heatmaps show the sensitivity of the first class for each binary classifier.

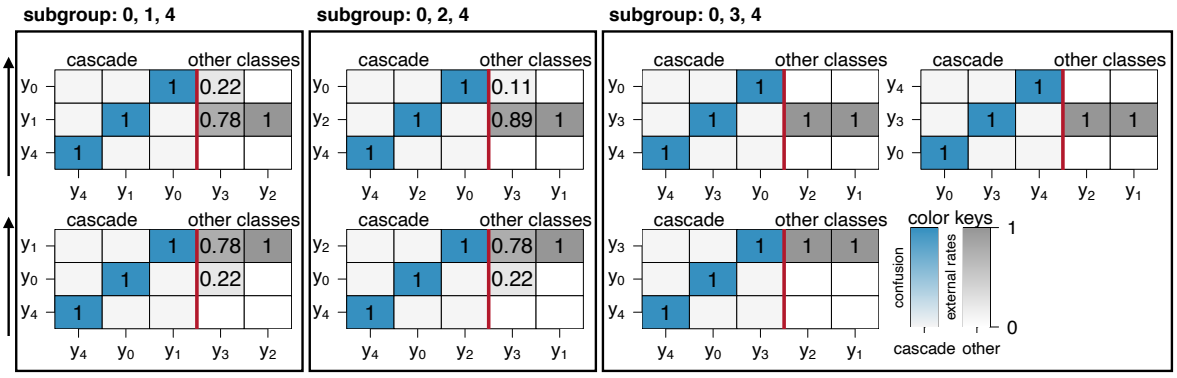


Figure A7: Confusion matrices on the returned subcascades for L1-support vector machine (SVM) homogenous ordinal classifier cascade (hOcc) on the pancreatic adenocarcinoma (PDAC) RNA-Sequencing dataset (d_{10}). The subcascades are grouped based on the classes used to build the cascades. The cascade classes are given in the rows and ordered from bottom to top. The columns comprise all classes and the values correspond to the confusion values indicating the amount of samples of the class given in the column that is assigned to the class given in the row.

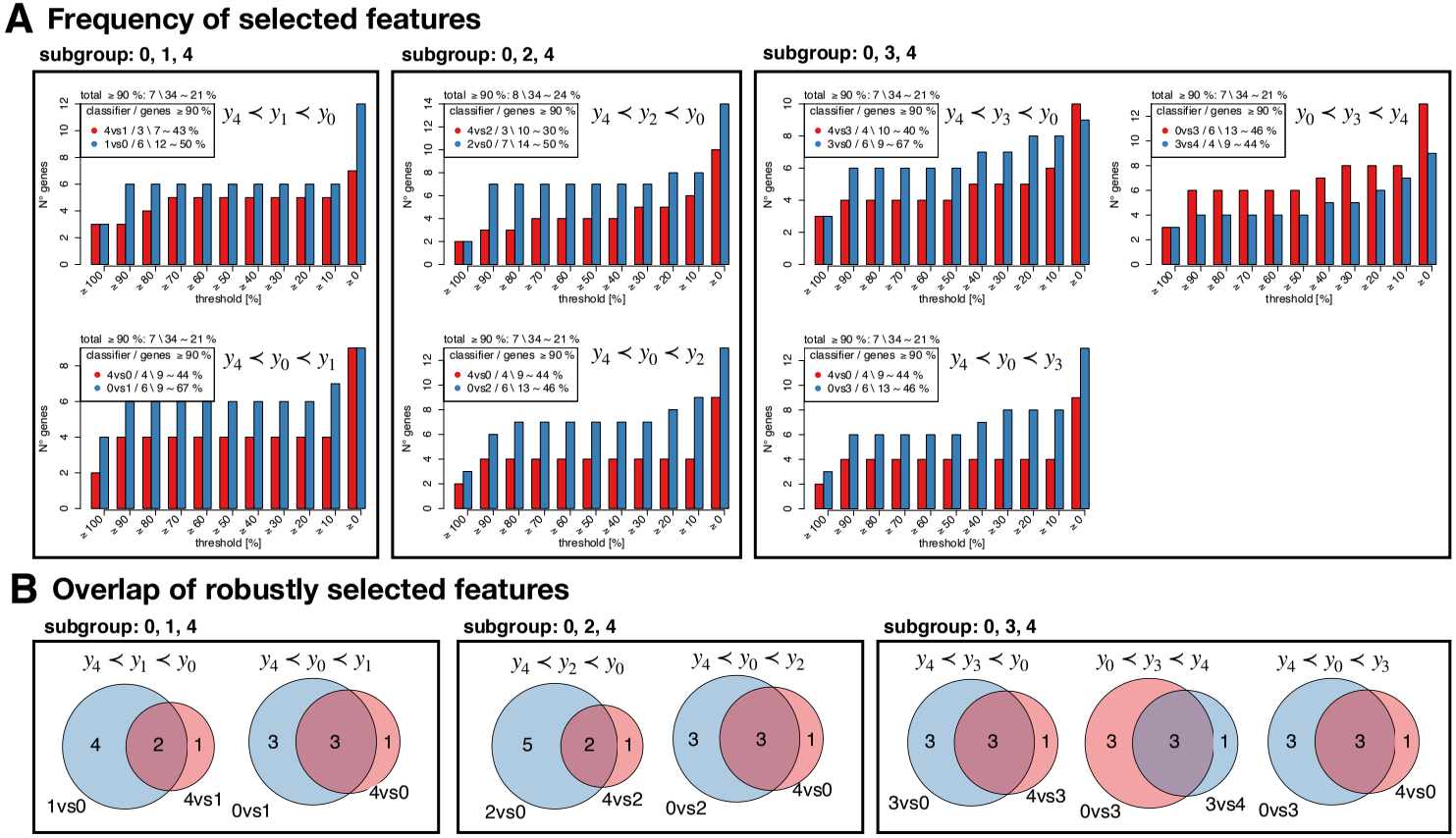


Figure A8: Results on the features for the returned subcascades for L1-support vector machine (SVM) homogenous ordinal classifier cascade (hOcc) on the pancreatic adenocarcinoma (PDAC) RNA-Sequencing dataset (d_{10}). The subcascades are grouped based on the classes used to build the cascades. A: Histogram on the selected features for each finally selected cascade. A threshold of 90 % is set to decide for robustly selected features. The total number of genes as well as the number and percentage of genes passing this threshold is noted at the top. The same measures are given for each base classifier in the legend. For the total number, features that are selected in more than one base classifier are counted only once. At least two permutations per group have been returned. B: Overlap of the selected features of the base classifiers that make up the finally selected cascades. The number of the robust features (selected in $\geqslant$ 90 % of the folds) per base classifier (colour coded) that make up the finally selected cascades.

Potential of correction methods for false positive edges (Section 6.2.2.2)

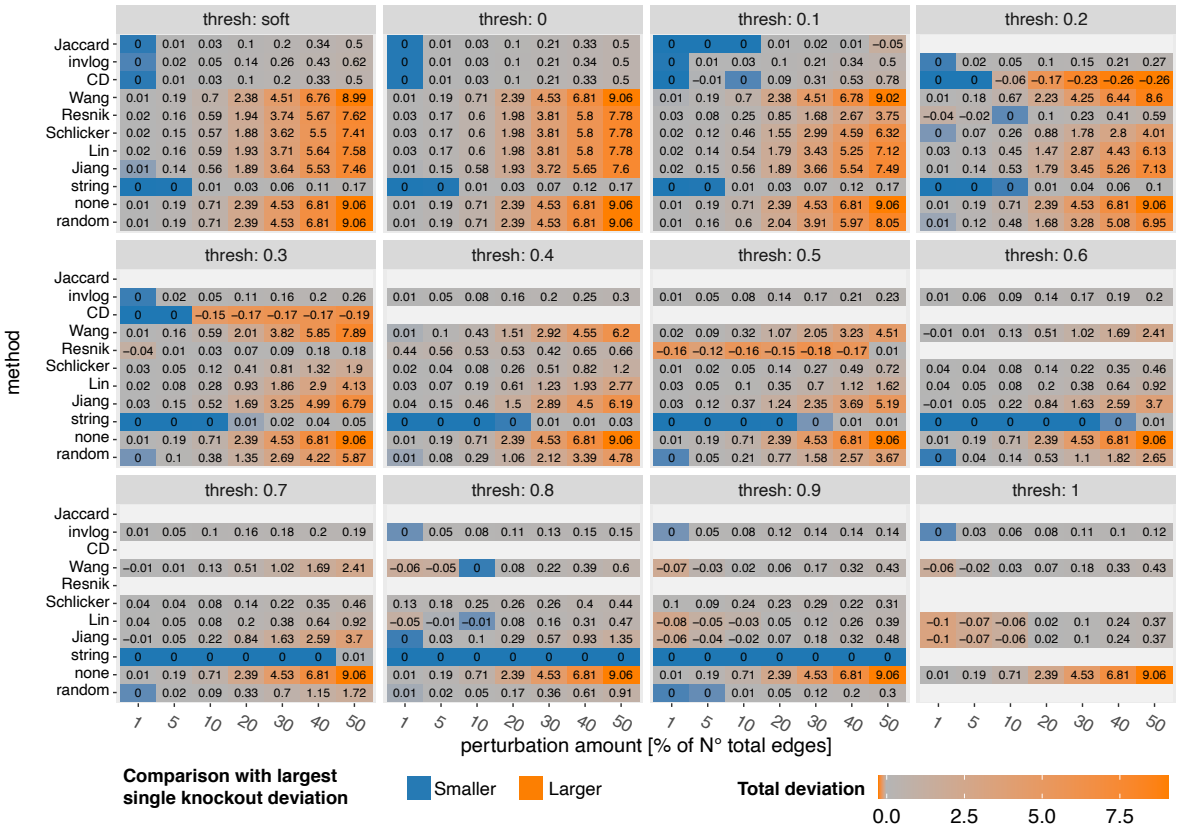


Figure A9: Performance of false positive correction techniques. The relative deviation between the local network entropy of the unperturbed corrected network and the perturbed corrected network is shown (stochastic deviation). Each panel corresponds to one threshold and comprises the various methods in the rows and the amount of perturbation in the columns. The colour coding is with respect to the maximal deviations measured between a single knockout and a wildtype stage. Blue colour corresponds to deviations smaller than the deviation between knockout and wildtype and orange to deviations larger than these thresholds.

Network entropy courses of (pre-)cancer stages (Section 6.3)

Table A5: P-value test statistics using Mann-Whitney-U test and Holm correction [89]. Denoted are the corrected p-values comparing the global entropy rates (global) and the local entropies (local) for the classes given in the column pairs for colon carcinoma (d_7), pancreatic adenocarcinoma (PDAC), Microarray (d_8), hepatocellular carcinoma (d_9), PDAC, RNA-Sequencing (d_{10}).

d_7: Colon carcinoma

pair	global	local	pair	global	local	pair	global	local
y_0 vs y_1	3.39E-05	4.77E-02	y_0 vs y_3	4.01E-06	2.97E-02	y_1 vs y_3	3.36E-02	4.51E-01
y_0 vs y_2	1.00E-07	1.20E-04	y_1 vs y_2	8.47E-05	6.48E-02	y_2 vs y_3	3.53E-01	4.51E-01

d_8: PDAC, Microarray

pair	global	local	pair	global	local	pair	global	local
y_0 vs y_1	1.00E+00	4.55E-01	y_1 vs y_2	1.00E+00	1.00E+00	y_2 vs y_4	1.97E-01	1.00E+00
y_0 vs y_2	1.80E-01	1.80E-01	y_1 vs y_3	1.80E-01	8.83E-01	y_3 vs y_4	1.00E+00	1.00E+00
y_0 vs y_3	1.00E+00	1.00E+00	y_1 vs y_4	1.00E+00	1.00E+00			
y_0 vs y_4	1.00E+00	4.66E-02	y_2 vs y_3	8.55E-02	2.74E-01			

d_9: Hepatocellular carcinoma

pair	global	local	pair	global	local	pair	global	local
y_0 vs y_1	1.00E+00	2.58E-01	y_1 vs y_8	1.87E-01	1.33E-03	y_4 vs y_5	1.00E+00	2.16E-02
y_0 vs y_2	1.00E+00	1.00E+00	y_1 vs y_9	2.75E-04	2.28E-02	y_4 vs y_6	1.00E+00	1.00E+00
y_0 vs y_3	2.96E-01	1.00E+00	y_2 vs y_3	1.00E+00	1.00E+00	y_4 vs y_7	1.00E+00	2.95E-02
y_0 vs y_4	1.00E+00	1.84E-01	y_2 vs y_4	1.00E+00	1.00E+00	y_4 vs y_8	2.13E-02	1.33E-04
y_0 vs y_5	1.00E+00	1.00E+00	y_2 vs y_5	1.00E+00	7.65E-02	y_4 vs y_9	2.42E-04	4.86E-03
y_0 vs y_6	1.00E+00	1.00E+00	y_2 vs y_6	1.00E+00	8.40E-01	y_5 vs y_6	1.00E+00	1.00E+00
y_0 vs y_7	1.00E+00	1.00E+00	y_2 vs y_7	1.00E+00	2.18E-01	y_5 vs y_7	1.00E+00	1.00E+00
y_0 vs y_8	2.87E-02	2.19E-03	y_2 vs y_8	5.16E-02	1.33E-04	y_5 vs y_8	2.30E-01	1.80E-01
y_0 vs y_9	2.87E-04	3.62E-01	y_2 vs y_9	1.21E-03	3.06E-02	y_5 vs y_9	1.23E-03	1.00E+00
y_1 vs y_2	1.00E+00	1.00E+00	y_3 vs y_4	2.30E-01	2.33E-01	y_6 vs y_7	1.00E+00	1.00E+00
y_1 vs y_3	9.11E-01	2.18E-01	y_3 vs y_5	1.00E+00	9.21E-01	y_6 vs y_8	1.00E+00	8.40E-01
y_1 vs y_4	1.00E+00	1.00E+00	y_3 vs y_6	1.00E+00	1.00E+00	y_6 vs y_9	4.45E-02	1.00E+00
y_1 vs y_5	1.00E+00	1.19E-01	y_3 vs y_7	1.00E+00	1.00E+00	y_7 vs y_8	1.00E+00	3.62E-01
y_1 vs y_6	1.00E+00	1.00E+00	y_3 vs y_8	6.15E-01	6.04E-04	y_7 vs y_9	1.52E-01	1.00E+00
y_1 vs y_7	1.00E+00	8.39E-02	y_3 vs y_9	2.11E-02	2.18E-01	y_8 vs y_9	1.00E+00	1.00E+00

d_{10}: PDAC, RNA-Sequencing

pair	global	local	pair	global	local	pair	global	local
y_0 vs y_1	1.00E+00	1.00E+00	y_1 vs y_2	1.00E+00	1.00E+00	y_2 vs y_4	1.00E+00	1.00E+00
y_0 vs y_2	1.00E+00	1.00E+00	y_1 vs y_3	1.00E+00	1.00E+00	y_3 vs y_4	1.00E+00	3.14E-02
y_0 vs y_3	1.00E+00	5.73E-01	y_1 vs y_4	1.00E+00	1.00E+00			
y_0 vs y_4	1.00E+00	1.00E+00	y_2 vs y_3	1.00E+00	1.00E+00			

STATUTORY DECLARATION

I hereby declare that I wrote the present dissertation with the topic:
System biology of tumour evolution: estimating orders from omics data
independently and used no other aids that those cited. In each individual case, I have clearly identified the source of the passages that are taken word for word or paraphrased from other works.
I also hereby declare that I have carried out my scientific work according to the principles of good scientific practice in accordance with the current „Satzung der Universität Ulm zur Sicherung guter wissenschaftlicher Praxis“ [Rules of the University of Ulm for Assuring Good Scientific Practice].

Ulm,

Lisa Monika Schäfer