# An MHN-Based Distance Measure for Clustering Cancer Progression Events

Bachelor's Thesis
of


Michael Bonart
2373397

University of Regensburg
Faculty of Informatics and Data Science
Department of Statistical Bioinformatics

Advisor: Prof. Dr. Rainer Spang


August 14, 2025

# Contents

# 1 Introduction

Understanding the biological signaling mechanisms in (human) cells is an ongoing active field of research, which has a wide range of applications in medicine, especially in cancer treatment. This knowledge is crucial for understanding how cancer diseases can be treated and is used to explain how tumors arise and evolve. Knowledge about cancer progression - the mechanisms by which healthy cells can turn into cancerous ones - thus relies hugely on our comprehension of cell signaling (Hancock (2017): Chapter 1). On the other hand our knowledge about a cell's biological signaling mechanisms might also greatly benefit from more in-depth observations on cancer progression, which is an idea that we will partly explore in this thesis.

In cancer progression a great number of genomic events (e.g. mutations, copy number variations and changes in DNA methylation and gene expression) are known to play an essential role. Many events' occurrences are observed to be stochastically dependent on each other (see e.g. Figs. 8 and 10), which is canonically explained by the darwinistic selection of tumor cells, whose accumulated genomic events yield them the most reproductive fitness. Whether a single genomic event benefits or harms a tumor cell's reproductive fitness may thereby also depend on other previously accumulated events (Nowell (1976)). These stochastic dependencies between genomic events' occurences can be interpreted in a multitude of ways, among which are cancer progression models like Mutual Hazard Networks (MHNs) introduced by Schill et al. (2019). MHNs model the occurrence of genomic events in cross-sectional data of cancer patients in a way that can be optimized to fit a given dataset with a maximum likelihood estimation. The parameters inferred from such an MHN optimization include influence factors between events that each translate to one event's occurrence hindering/facilitating the accumulation of another event. In this thesis we explore the possibility of constructing a distance measure for genomic events based on these influence factors. In our distance measure two events shall be considered close if they play interchangeable roles in the web of influences between events in cancer progression. In the context of MHN this would mean having similar influence factors to the same events. Such interchangeability in cancer progression could be used to cluster cancer progression events, better understand their relevance and also hint at related roles of the affected proteins in biological signaling pathways.

All implementations used in this thesis are publicly available online (see Appendix A).

# 2 Distance Measure for interchangeability of cancer progression events

Our proposed distance measure for interchangeability of cancer progression events is always based on a dataset $\mathscr{D}$, which is a binary matrix containing the occurrence patterns of a set of genomic events $E$ across a group of $n \in \mathbb{N}$ cancer patients.

$$\mathscr{D} = (d_i)_{i=1..n} \qquad\qquad d_i \in \{0,1\}^E$$

## 2.1 MHN: Mutual Hazard Networks

The main tool we use to extract relevant information from our dataset is MHN, which is a parametrized cancer progression model first introduced by Schill et al. (2019).
For our purposes we will regard MHN as a function $MHN : \{0,1\}^{n\times E} \to \mathbb{R}^{E\times E}$, where $MHN(\mathscr{D}) = \theta$ is the result of a cMHN-optimization on the dataset $\mathscr{D}$ as defined in the publicly available mhn-package (see `https://pypi.org/project/mhn`).

The parameters in $\theta \in \mathbb{R}^{E\times E}$ can be interpreted as a high-likelihood explanation for mutual influences between events in $E$ during their accumulation process:

- An event's base rate (its diagonal entry $\theta_{ii}$) defines the base probability of the event occurring.
- Each offdiagonal entry $\theta_{ij}(i \neq j)$ defines how the probability of event $i$ occurring changes if event $j$ is already present. These influences between genomic events are attributed to their combined influences on the reproductive fitness of a tumor cell.

More precisely, MHN specifies a continuous time Markov process, in which the transition rate $Q_{\mathbf{x}_{+i},\mathbf{x}}$ for the transition from $\mathbf{x} = (...,x_{i-1},0,x_{i+1},...)^T$ to $\mathbf{x}_{+i} := (...,x_{i-1},1,x_{i+1},...)^T$ (accumulation of event $i$) is given by:

$$Q_{\mathbf{x}_{+i},\mathbf{x}} = exp\left(\theta_{ii} + \sum_{i\neq j} \theta_{ij}x_j\right)$$

## 2.2 Reduced MHNs

To define our distance measure, we split the set of available events $E$ into two subsets: The set of test events $T$, which will determine our distance measure, and the set of cluster events $\Omega$, whose distances from each other we want to measure. For our algorithm it is necessary that these two subsets are disjunct $\Omega \cap T = \emptyset$, usually we choose $\Omega = E \setminus T$ with $T$ containing about 2 - 6 events. The distance between two events in $\Omega$ is computed by comparing their interactions with the test events.

The interactions of an event $e \in \Omega$ with the test events are determined by training an MHN on a reduced projection of the dataset that contains only data for events $T \cup \{e\} \subseteq E$.
For $M \subseteq E$ and the canonical projection $pr_M : \{0,1\}^E \to \{0,1\}^M$ we define

$$\widehat{\mathscr{D}}_M = (\widehat{d_i})_{i=1..n} \qquad \widehat{d_i} = pr_M(d_i) \in \{0,1\}^M$$

as the onto $M$ reduced dataset.
On these reduced datasets $\widehat{\mathscr{D}}_{T\cup\{e\}}$ it is computationally feasible and even fast to train many MHNs, which we use as the basis for distance calculation between two events $e_1, e_2 \in \Omega$:

$$dist_{\mathscr{D},T}(e_1,e_2) := dist(\theta_{e_1}, \theta_{e_2}) \qquad \theta_e = MHN(\widehat{\mathscr{D}}_{T\cup\{e\}})$$

Note that this distance measure will always depend on the given dataset $\mathscr{D}$, the choice of test events $T$ and the details of MHN optimization.

## 2.3 MHN regularization

For all instances of MHN optimization in this thesis we regularize $\theta$ by imposing an L1 penalty on the offdiagonals. It is always weighted by the factor $\lambda = \frac{1}{|\mathscr{D}|}$ to ensure good comparability between MHNs trained on the same dataset $\mathscr{D}$. The more elaborate approach of determining $\lambda$ in cross-validation was also explored, which confirmed that our choice for $\lambda$ lies within a reasonable interval close to the optimal value.

## 2.4 MHN identifiability and comparison

For datasets with few events ($|E| \leq 4$) it has been shown that unregularized MHN optimization suffers from identifiability issues, where different parameters $\theta$ induce the same probability distributions (Vocht (2022)). As we make use of an L1-regularization we are not directly affected by this, however we are concerned with a related phenomenon, where very similar distributions can lead to very different $\theta$'s.

The most commonly observed ambiguities for small MHNs were stochastic dependencies between events $i, j$ being depicted either as an influence of $i$ on $j$ or vice versa as an influence of $j$ on $i$. In MHN this shows up as offdiagonal entries $\theta_{ji}$ or $\theta_{ij}$. We combat these ambiguities by symmetrizing $\theta$'s before comparing them using a L1-distance. Additionally we choose to only compare the offdiagonal entries as our focus is on interactions with the test events. The base rates (diagonal entries) are therefore ignored, ensuring that our distance measure is able to cluster together events of differing absolute frequencies too.

$$dist(\theta_1,\theta_2) := \sum_{i\neq j} |sym(\theta_1)_{ij} - sym(\theta_2)_{ij}| \qquad sym(\theta) = \frac{\theta + \theta^T}{2}$$

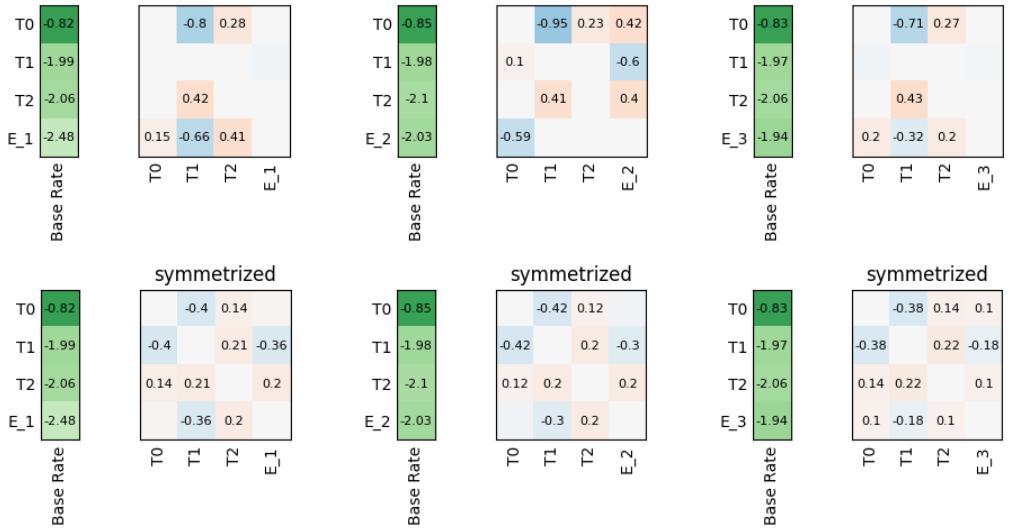



Figure 1: Examples for three $\theta$'s with identical test events and differing cluster events $E_1, E_2$ or $E_3$, which are similarly distributed. The $\theta$-matrices show significant differences, while the symmetrized versions reveal a more close relation. (Blank offdiagonal spaces are entries close to zero $|\theta_{ij}| < 0.1$; base rates are displayed in a separate column to the left instead of on the diagonal)

## 2.5 Concise mathematical formulation

The distance measure proposed above can be written as a composition of an MHN optimization and two linear projections, whose image is compared by an L1-distance. It depends on $\mathscr{D} \in \{0,1\}^{n \times E}$ and $T \subseteq E$:

$$dist_{\mathscr{D},T} : \Omega \times \Omega \to \mathbb{R}_0^+ \qquad (e_1, e_2) \mapsto \sum_{i \neq j} \left| (\theta_{T,e_1})_{ij} - (\theta_{T,e_2})_{ij} \right|$$

$$\text{with } \Omega = E \setminus T \qquad \theta_{T,e} = \left( sym \circ MHN \circ pr_{T \cup \{e\}} \right) (\mathscr{D})$$

Note: Although $\theta_{T,e}$ refers to the used $\theta$-matrices after symmetrization (see definition above), displays of $\theta_{T,e}$'s in this thesis will usually be in their not yet symmetrized form to accurately reflect the results of the underlying MHN optimizations.

## 2.6 Normalization and combined distance measures

Two key shortcomings of the above $dist_{\mathscr{D},T}$ are its limited domain $\Omega$, which doesn't include all available events, and a high dependence on which test event sets $T$ are used as different choices of test events sets can lead to very different clusterings. To construct a more stable distance measure, whose domain is all of $E$, we combine multiple distance measures. As these distance measures are potentially based on $\theta_{T,e}$'s of varied sizes we normalize each distance measure $dist_{\mathscr{D},T}$ with a factor of $\frac{1}{|T|}$. This accounts for the number of entries that we expect to differ in two $\theta_{T,e} \in \mathbb{R}^{(T \cup \{e\}) \times (T \cup \{e\})}$ for different $e$.

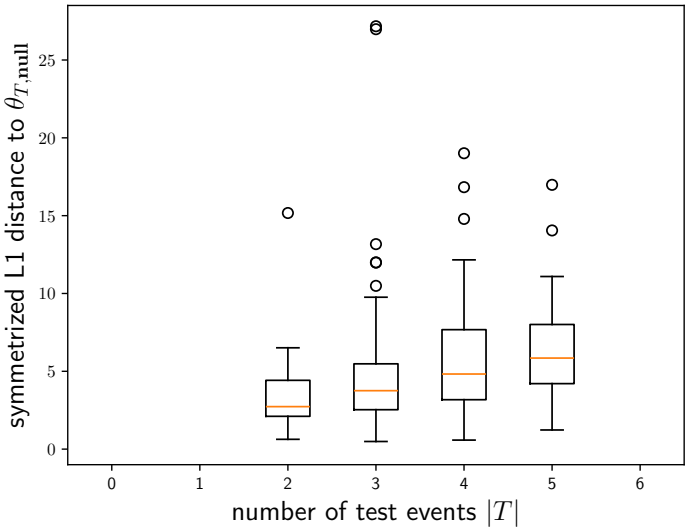



Figure 2: The median L1 distances between symmetrized $\theta_{T,e}$'s increase linearly with the number of test events. $\theta_{T,\textbf{null}}$ is an MHN trained on only the test events $T$ with an added zero row and column. It corresponds to an event that has no interactions with the test events.

For a set of test event sets $\mathbb{T} \subseteq \mathscr{P}(E)$ with $|\mathbb{T}| \geq 3$ and a dataset $\mathscr{D} \in \{0,1\}^{n \times E}$ we define the combined distance measure:

$$\overline{dist}_{\mathscr{D},\mathbb{T}} : E \times E \to \mathbb{R}_0^+ \qquad (e_1, e_2) \mapsto avg\left(\left\{ \frac{dist_{\mathscr{D},T}(e_1,e_2)}{|T|} \text{ for } T \in \mathbb{T} \text{ with } e_1, e_2 \notin T \right\}\right)$$

(with *avg* being the arithmetic mean)

Note: To have a well defined distance for every pair $(e_1, e_2) \in E \times E$, at least 3 test event sets are required. If there are only 3 test events sets, these additionally have to be disjunct.

# 3 Evaluation on simulated data

To evaluate the clustering capabilities of our distance measure $dist_{\mathscr{D},T}$, we generate ground truth MHNs with intended known clusterings and sample data from them. Based on the sampled data we try to recover the intended clustering using our distance measure.

## 3.1 Generate MHNs and datasets with ground truth clustering

We first arbitrarily generate an MHN containing test events $T = \{T0, T1, T2, \dots\}$ and event clusters $\Omega_{/\sim} = \{A, B, C, \dots\}$, which constitute the intended clustering for cluster events $\Omega = \{A_0, A_1, A_2, B_0, B_1, \dots\}$ (with $A = \{A_0, A_1, A_2\}$, $B = \{B_0, B_1\}, \dots$). To ensure good convergence properties, we sample data from the arbitrary MHN and use it to train a new MHN, with which we will continue working. Such a training iteration may be repeated multiple times.

Once we have a satisfying ground truth MHN $\theta \in \mathbb{R}^{(T \cup \Omega_{/\sim}) \times (T \cup \Omega_{/\sim})}$ we split up each event cluster $e \in \Omega_{/\sim}$ into its events $e_i \in \Omega$ $(i = 0, ..., |e| - 1)$ that are supposed to form a cluster. The split MHN's offdiagonal values for $e_i$ are copied from the original MHN's values for $e$, while the base rates (diagonal) are replaced by independently chosen random values.

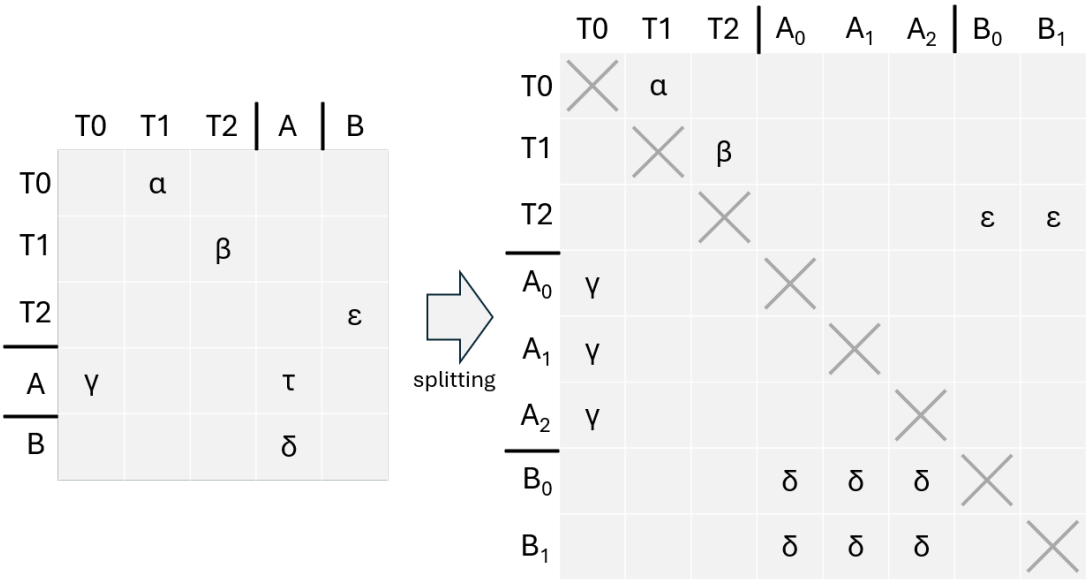



Figure 3: Example for splitting up event clusters in the ground truth MHN. Values from the left-hand MHN are copied to the right-hand MHN. Base rates (such as $\tau$ in this example) are ignored and replaced by random values.

For our evaluation we will try to recover the given clustering solely based on data sampled from the split MHN. Sampling is done via its python implementation in the mhn-package, where $sample_n : \mathbb{R}^{E \times E} \to \{0,1\}^{n \times E}$ maps $\theta$ to a simulated dataset $\mathscr{D}$ of size $|\mathscr{D}| = n$, which fits the distribution defined by $\theta$, such that for large $n \to \infty$ we approach the relationship $MHN \circ sample_n = id_{Im(MHN)}$.

## 3.2 Adding perturbation and noise

For evaluation we will add perturbation and noise in our data generation process. The noise comes from our sampling process, which is always influenced by random effects as we only generate finitely sized datasets. This lets us estimate how large a dataset needs to be in order to make meaningful statements about two events' similarity. The perturbation affects the ground truth distribution. We multiply the split ground truth MHN's offdiagonal entries by samples from a gaussian distribution centered around 1. This lets us estimate how similar two events need to be in the ground truth MHN to still be properly clustered together using our algorithm. For a perturbation level $\sigma \in \mathbb{R}_0^+$ and a given MHN $\theta \in \mathbb{R}^{E \times E}$ we define $\widehat{\theta}_\sigma$ as a randomly generated MHN close to $\theta$, whose values are sampled as follows ($\mathcal{N}(\mu, \sigma^2)$ is a gaussian distribution):

$$(\widehat{\theta}_\sigma)_{ij} \sim \begin{cases} \mathcal{N}(1, \sigma^2) \cdot \theta_{ij} & \text{if } i \neq j \\ \theta_{ij} & \text{if } i = j \end{cases}$$

(base rates are not altered as they are already completely randomized)

The datasets generated from the above procedure will be referred to as $\mathcal{D}_{n,\sigma}$.

$$\mathcal{D}_{n,\sigma} := sample_n(\widehat{\theta}_\sigma)$$

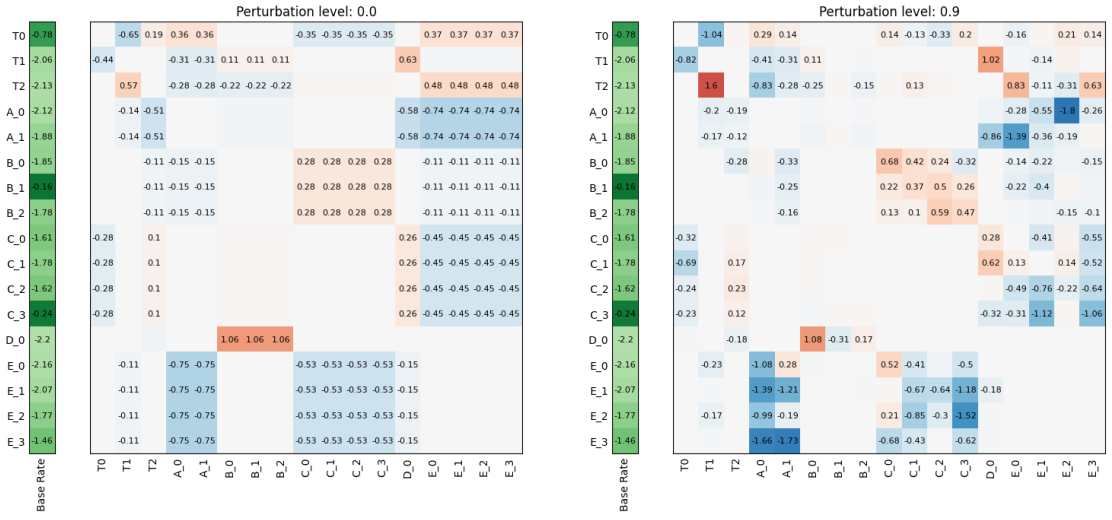



Figure 4: Example of two ground truth MHNs $\widehat{\theta}_{\sigma=0.0}, \widehat{\theta}_{\sigma=0.9}$ with different perturbation levels on the same underlying $\theta$.

For one ground truth $\theta$, whose correct clustering is known, and a predefined set of test events $T = \{T0, T1, T2\}$ we compute the distance measure $dist_{\mathcal{D}_{n,\sigma},T}$ for different perturbation levels $\sigma \in \{0.0, \ldots, 0.9\}$ and sample sizes $n \in \{1000, \ldots, 100000\}$. Based on the resulting distance matrices we compute hierarchical clusterings, which we cut into clusters to compare them to the known ground truth clustering using the adjusted rand index.

The adjusted rand index $ARI(\sim_1, \sim_2) \in [-1, 1]$ introduced by Hubert and Arabie (1985) compares two partitionings $\sim_1, \sim_2$ of a set and returns 1 for identical partitionings and 0 for unrelated partitionings. To infer a partitioning from a hierarchical clustering, we cut it at one height level determined by the desired number of resulting clusters $n_{cluster}$. To make an $ARI$ of 1 achievable, $n_{cluster}$ is chosen to match the number of clusters in the ground truth clustering. The adjusted rand index of these partitionings are our main metric for scoring and evaluating the underlying distance measures. As an alternative metric we also show the maximum adjusted rand index that is achievable when cutting the hierarchical clustering at any one height level.

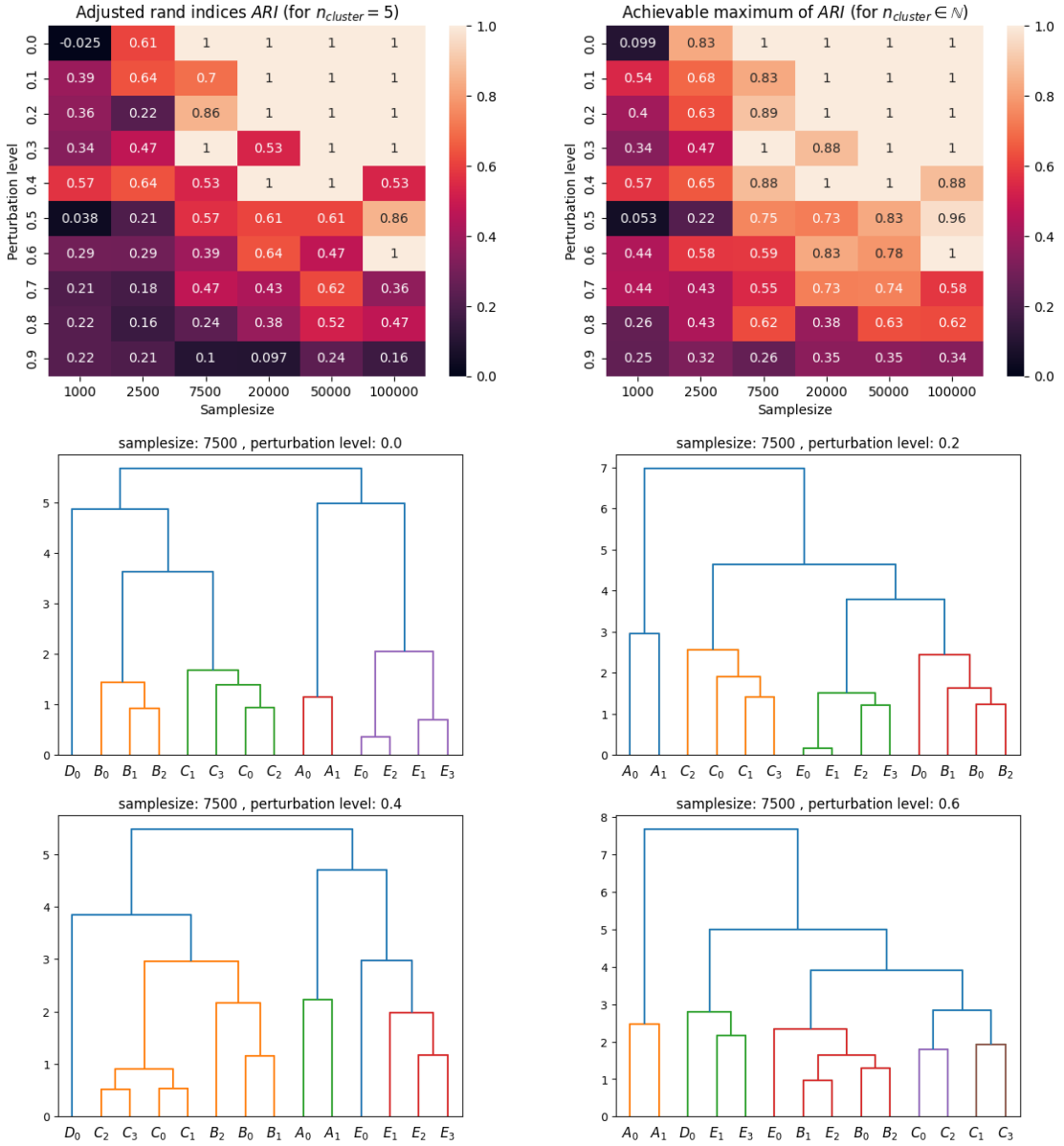



Figure 5: Top: Adjusted rand indices for different perturbation levels and sample sizes. (left: correct number of clusters was given; right: maximum possible adjusted rand index when cutting at any one height level)
Bottom: Four examples of hierarchical clusterings, whose adjusted rand indices are shown above. (colors indicate partitioning for $n_{cluster} = 5$)

## 3.3 Comparison to non-symmetrizing distance

Ignoring the identifiability arguments discussed in 2.4, one might also choose to compare the $\theta$'s of two events more simply without symmetrizing:

$$dist_{non-sym}(\theta_1, \theta_2) := \sum_{i \neq j} \left| (\theta_1)_{ij} - (\theta_2)_{ij} \right|$$

Evaluation results using this alternative approach reveal however that this non-symmetrizing L1-distance is unsuccessful in finding the correct clusters. Thus symmetrization is infact an important part of our distance measure.

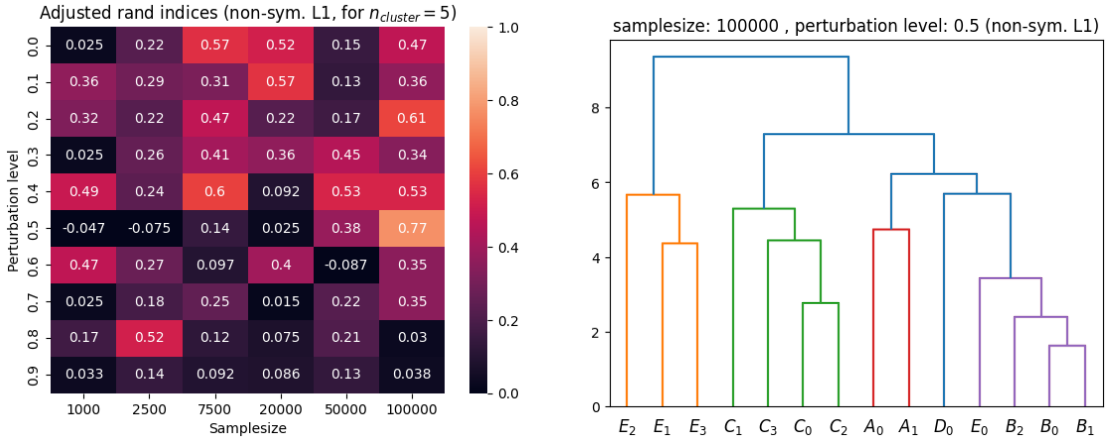



Figure 6: For the non-symmetrizing L1-distance the best achieved adjusted rand index in this evaluation was 0.77. A completely correct clustering was found in none of the cases. On the right-hand side we display the linkage that achieved an adjusted rand index of 0.77.

# 4 Application on selected datasets

We apply the combined distance measure (as defined in 2.6) on cross-sectional data of patients suffering from selected types of cancer. By this method we try to identify clusters of genomic events with interchangeable roles in the progression of these cancer types. The genomic events we consider for clustering are mutations (M) and copy number variations including deletions (Del) and amplifications (Amp). We additionally suggest that some clusters of genomic events could hint at related biological signaling roles of the proteins encoded on the events' underlying genes.

## 4.1 Link to biological pathways

Biological signaling pathways (biopathways) are molecular reaction chains within a cell that trigger and determine its actions including its metabolism, regulation of gene expression and transduction of signals (Hancock (2017): Chapter 1.2). Among the crucial molecules making up a biopathway are proteins encoded for in genes. Therefore mutations on these protein-encoding genes (which can be measured and observed using sequencing methods (Sinclair (2002))) directly affect the pathways that their proteins are part of. This can alter the cell's behavior, change its reproductive fitness and potentially turn it or enable it to turn cancerous (Hancock (2017): Chapter 1.10). Though this basic mechanism is well established, many details of the molecular interactions within cells are not fully understood.

Our newly introduced distance measure primarily aims to improve our understanding of cancer progression. However we additionally try to extend its applicability to biopathways and molecular interactions within the cell by finding clusters of proteins with similar roles in the cellular reaction mechanism. We relate such similarity of proteins' roles to genomic events (i.e. mutations and copy number variations) on their encoding genes playing interchangeable parts in cancer progression.

The rationale for establishing this relationship can be summed up as follows: Statistically speaking, whether a genomic event becomes manifest on a gene in a tumor cell, mostly depends on whether it improves the tumor cell's reproductive fitness or not (Nowell (1976)). Whether a cell becomes more fit by gaining said genomic event may for this reason also depend on other circumstances in the cell, which in turn are in large parts influenced by the cell's genes and their mutation/expression status. If two proteins play interchangeable or similar roles in a biopathway, mutations on their encoding genes will also have similar effects on the circumstances in the cell, which in turn affects the probability of other genomic events occurring. Such mutual influences between genomic events can be statistically modeled using MHN for example. As argued above we would expect proteins with similar/interchangeable functions as well as their related genomic events with similar effects on the cell's circumstances to consequently also share similar interaction rates in $\theta$-matrices of MHNs trained on data of cancer patients. These similar interaction rates are what our proposed distance measure aims to detect by the methods described in chapter 2.

## 4.2 Dataset on lung cancer patients

We analyze the dataset 'G16_CH_LUAD' (provided by A. Loesch) that contains cross-sectional data on 2585 lung cancer patients. We use the following 4 test event sets of biologically related genomic events (for more details on these choices see Appendix B):

$$\begin{aligned}\mathbb{T}_{LUAD} = \{ \quad & \{\text{STK11 (M), KEAP1 (M)}\}, \quad \{\text{FAT1 (M), APC (M)}\}, \\ & \{\text{TP53 (M), RB1 (M), RB1/13q (Del)}\}, \\ & \{\text{EGFR (M), MET (M), KRAS (M), BRAF (M)}\} \qquad \}\end{aligned}$$

### 4.2.1 Results

Computing our distance measure with the parameters above yields the following results:

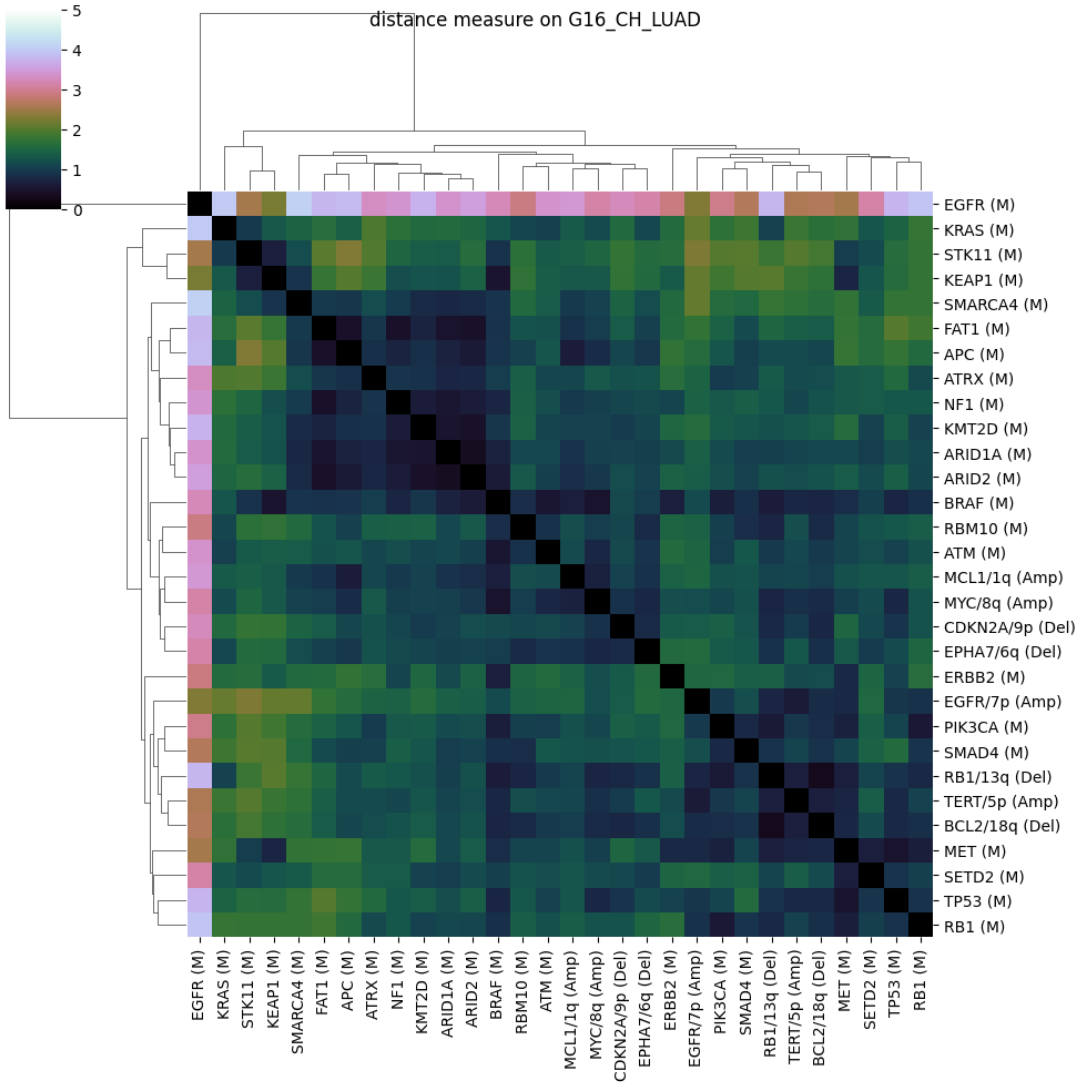



Figure 7: Combined distance measure $\overline{dist}_{\mathscr{D},\mathbb{T}_{LUAD}}$ with its inferred hierarchical clustering.

To better interpret the resulting distance matrix and its clustering, we additionally compare it to the correlation coefficients between each pair of events.

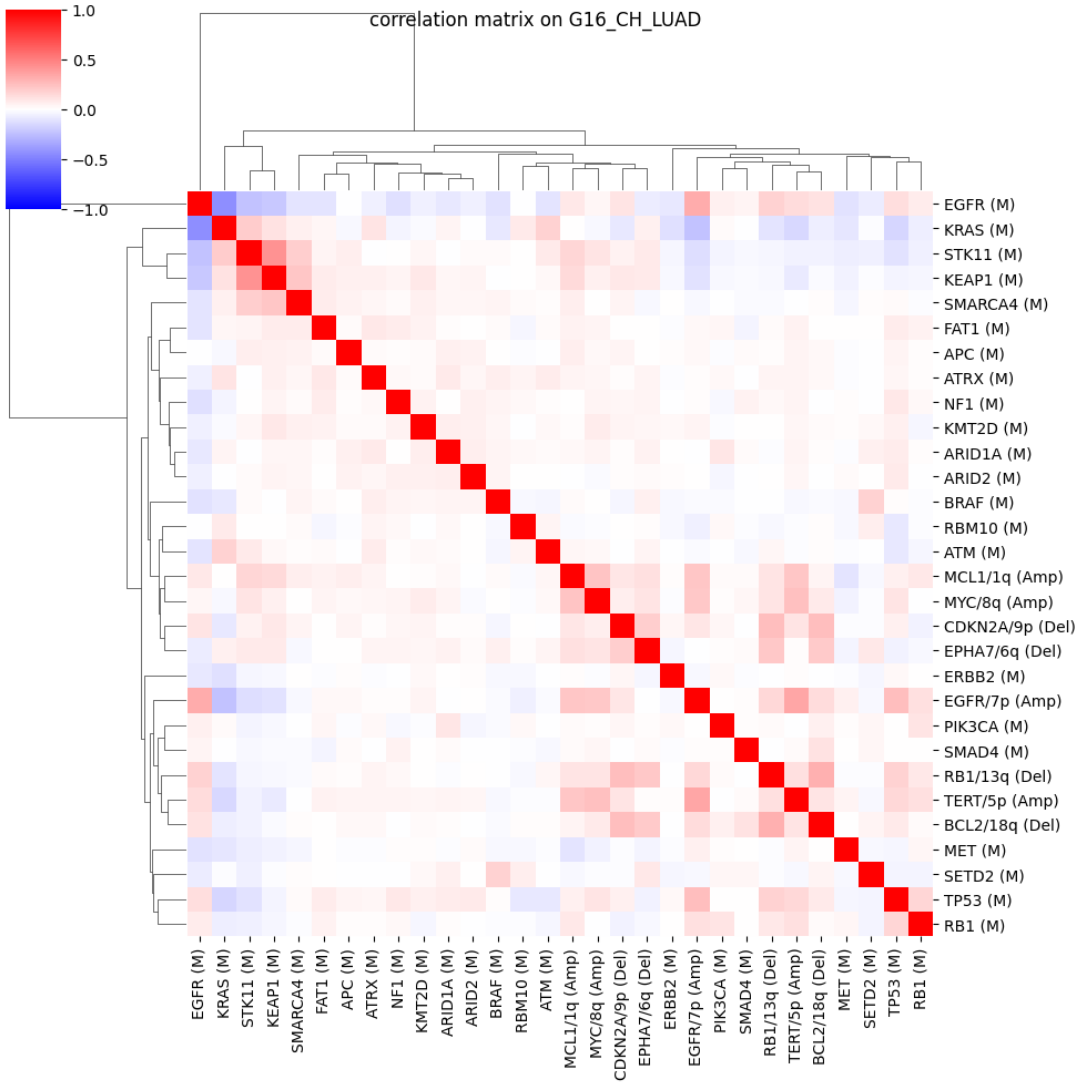



Figure 8: Correlations of events in G16_CH_LUAD with clustering from $\overline{dist}_{\mathscr{D},\mathbb{T}_{LUAD}}$.

### 4.2.2 Interpretation

We compare selected identified event clusters to existing findings about pathways and interactions concerning these genomic events. For event clusters that can not solely be explained by strong correlations we may additionally display some of their $\theta_{T,e}$ matrices. All displays of $\theta_{T,e}$'s in this chapter will be in their not yet symmetrized form to accurately reflect each MHN optimization's result.

ARID1A (M) / ARID2 (M):

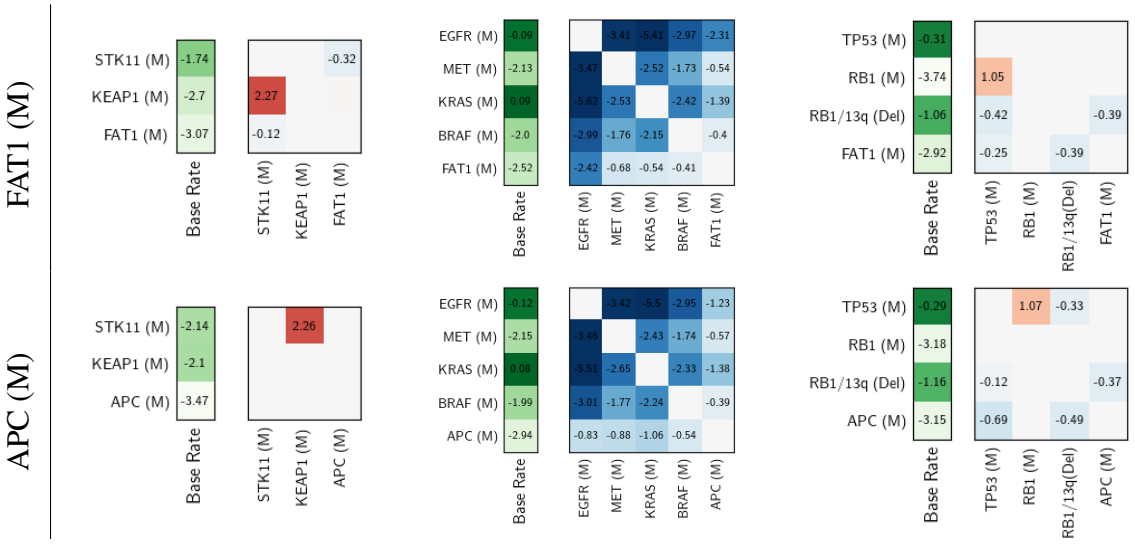
The proteins encoded in ARID1A and ARID2 are proven to take on equivalent roles in the homolog complexes BAF and PBAF (Hodges et al. (2016)). This is a succesful example of clustering together two mutation events on genes with interchangeable functionality.

STK11 (M) / KEAP1 (M):

Mutations of STK11 and KEAP1 are known to be highly correlated (see Fig. 8) in lung cancer patients and both individually promote ferroptosis protection, with their effects becoming most evident when they appear concurrently (Wohlhieter et al. (2020)). This is a good example of our distance measure finding a cluster of biologically related mutation events.

FAT1 (M) / APC (M):

FAT1 and APC have both been shown to promote Wnt-signaling when mutated (Morris et al. (2013); Sansom et al. (2004)). They were clustered together without being strongly correlated (see Fig. 8). Their symmetrized $\theta_{T,e}$'s are qualitatively similar with differences in the severity of some interactions (especially for EGFR (M)).




RBM10 (M) / ATM (M):

One major tumor suppressing mechanism of both RBM10 and ATM is inducing apoptosis by activating p53 (Jung et al. (2020); Cremona and Behrens (2014)). This could very well be the deciding explanation for their mutation events' closeness in our distance measure. While these two mutation events are hardly correlated with each other, they both are anti-correlated with TP53 (M) - the mutation event on p53's encoding gene (see Fig. 8). This further supports the argument that the hindering of the mechanism mentioned above is the essential consequence of RBM10 and ATM mutations in lung cancer progression.

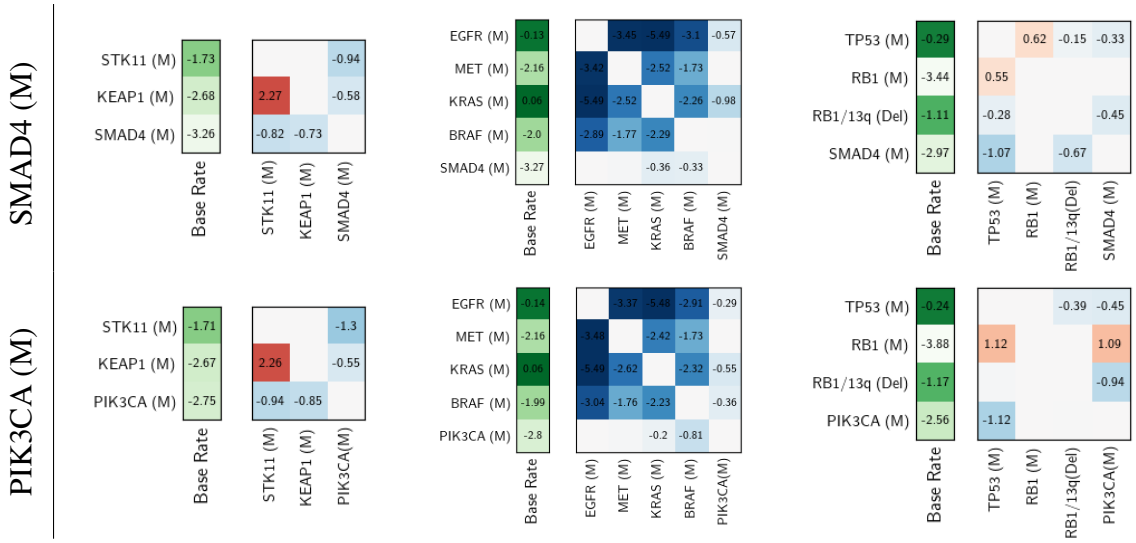
MCL1/1q (Amp) / MYC/8q (Amp):

Amplifications of MCL1 on chromosome 1q and amplifications of MYC on chromosome 8q were highly correlated in the chosen dataset (see Fig. 8). The overexpression of these genes has also previously been found to be related, which was explained by MCL1's ability to avert MYC-induced apoptosis (Allen et al. (2011)). For tumor cells this makes overexpression of MCL1 a very useful adaptive trait to survive while overexpressing MYC. Thus

these two genomic events could infact be considered interchangeable as their joint role in cancer progression is to enable overexpression of MYC.

SMAD4 (M) / PIK3CA (M):

SMAD4 and PIK3CA are commonly allocated to separate pathways (TGF-$\beta$/SMAD4 and PI3K/AKT). These could however be related through AKT's downstream effects involving the mTor-pathway that suppresses SMAD3 phosphorylation (Zhao et al. (2018)). In this case, it remains unclear whether a biologically related function is the reason for clustering these events together. The symmetrized $\theta_{T,e}$'s share similarities, but they have a substantial qualititative difference concerning the interactions with RB1 (M).




Expected groups not identified by the distance measure:

We also want to mention a negative example of a group of events that we expected to be clustered together based on previous findings that was however not identified as interchangeable by our distance measure.

KRAS (M) and BRAF (M) are mutations on genes, whose encoded proteins K-Ras and B-Raf are crucial components of the Ras/Raf/MAPK-pathway, which is responsible for the transduction of signals from the cell's microenvironment to its nucleus. As such B-Raf is one of the main targets of K-Ras, however there are other downstream targets of K-Ras like the PI3K cell survival pathway too (Molina and Adjei (2006)).

Our distance measure $\overline{dist}_{\mathscr{D},\mathbb{T}_{LUAD}}$ did not cluster these two events together, thus stating that their roles in cancer progression are not interchangeable. This could either stress the relevance of K-Ras' other downstream effects or generally hint at more molecular interactions relevant to K-Ras and B-Raf that allow for a clear distinction between them. The similarities and dissimilarities between these two mutations' effects on cancer progression and whether they can be considered interchangeable has also previously already been a topic of discussion (Oikonomou et al. (2014)).

## 4.3 Dataset on colorectal cancer patients

We analyze the dataset 'G16_CH_COAD' (provided by A. Loesch) that contains cross-sectional data on 1697 colorectal cancer patients. We use the following 5 test event sets of biologically related genomic events (for more details on these choices see Appendix B):

$$\begin{aligned}\mathbb{T}_{COAD} = \{ \quad & \{\text{KRAS (M), BRAF (M)}\}, \quad \{\text{ATM (M), TP53 (M)}\}, \\ & \{\text{TGFBR2 (M), SMAD4 (M)}\}, \\ & \{\text{FAT1 (M), APC (M), CTNNB1 (M), AMER1 (M)}\}, \\ & \{\text{ARID1A (M), KMT2D (M), KMT2C (M)}\} \qquad\qquad \}\end{aligned}$$

### 4.3.1 Results

Computing our distance measure with the parameters above yields the following results:

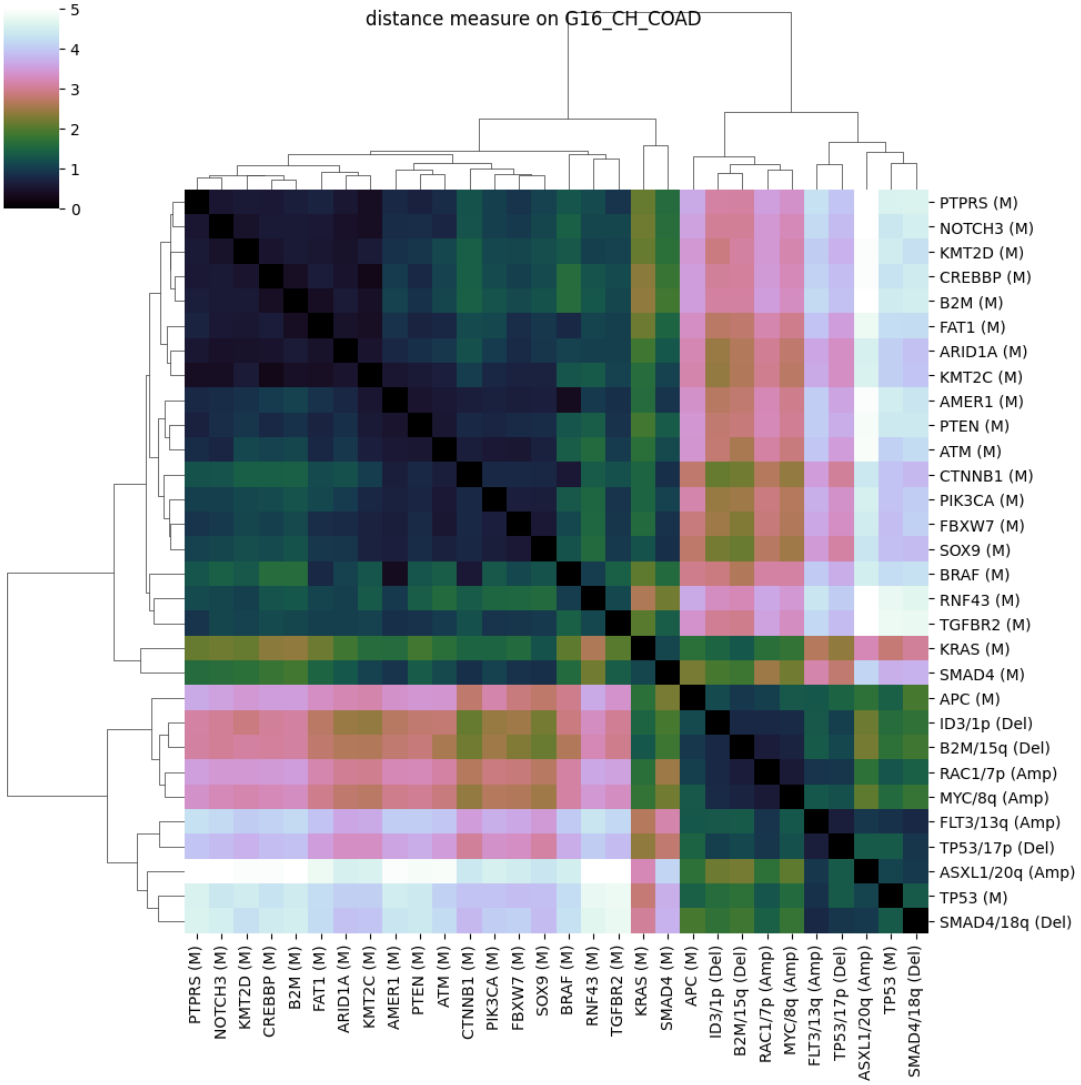



Figure 9: Combined distance measure $\overline{dist}_{\mathscr{D},\mathbb{T}_{COAD}}$ with its inferred hierarchical clustering.

To better interpret the resulting distance matrix and its clustering, we additionally compare it to the correlation coefficients between each pair of events.

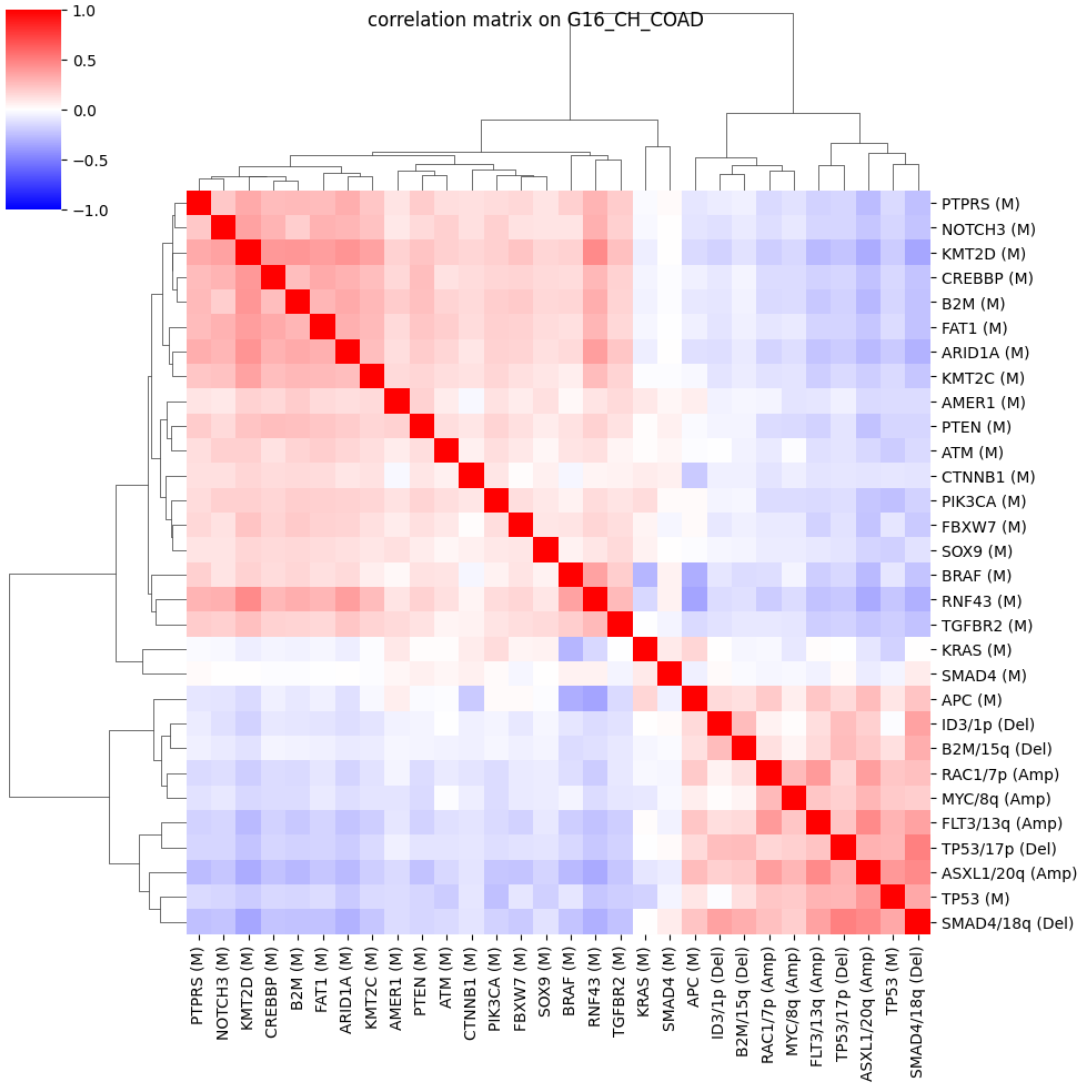



Figure 10: Correlations of events in G16_CH_COAD with clustering from $\overline{dist}_{\mathscr{D},\mathbb{T}_{COAD}}$.

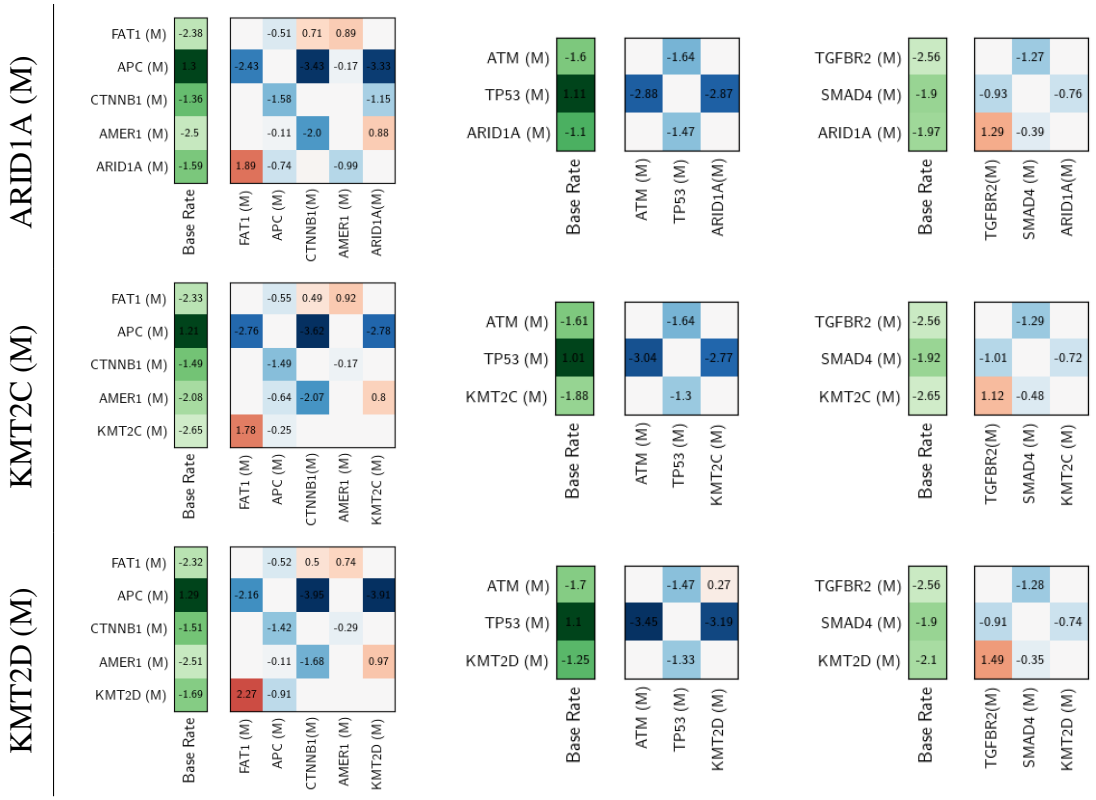
### 4.3.2 Interpretation

The prominent top-level divide of genomic events in Figure 9 clustered together all copy number variation events (with APC (M) and TP53 (M) being in their cluster too). This division is also clearly visible in the correlation matrix of the events (see Fig. 10) revealing two clusters of correlated events that are anticorrelated with the other cluster's events. This falls in line with a previously observed dichotomy of cancer types dominated by either mutation events or copy-number-variation events (Ciriello et al. (2013)). That the mutation events TP53 (M) and APC (M) are part of the copy-number-variation events' cluster can be explained by their shared property of causing chromosomal instability leading to the aforementioned copy number variations (Bronder et al. (2021); Fodde et al. (2001)).

We expect MHN's influence factors to be most telling for events occurring within the same cancer type as the test events, which is why our further interpretation will treat the copy number variations as one large cluster and from now on mostly focus on the subdivisions of the mutation events' cluster (the majority of used test events is in this cluster). Among these mutation events we identify smaller clusters and compare them to existing findings about pathways and interactions concerning these mutations.

ARID1A (M) / KMT2C (M):

ARID1A as part of the SWI/SNF-complex is responsible for chromatin remodeling, which has effects on DNA accessibility and gene expression (Wu et al. (2014)). KMT2C (M) was expected to be clustered with KMT2D (M) as KMT2C and KMT2D have lots of functional commonalities and are both used for methylation of histone H3K4, which marks enhancers and promoters (Froimchuk et al. (2017)). As such KMT2D and KMT2C also play an important role in shaping chromatin and regulating gene expression, which would explain the suggested interchangeability of ARID1A (M) and KMT2C (M) in cancer progression. Why KMT2D (M) was not part of this cluster remains unclear as of now, but could potentially hint at subtle differences between KMT2D (M) and KMT2C (M). Although their distance was not necessarily large, there were many other events close to both of them in the distance matrix. Below we display some $\theta_{T,e}$'s of the events mentioned above to better illustrate their similarities and dissimilarities.




CREBBP (M) / B2M (M):

B2M is part of the major histocompatibility complex class I (MHC-I) that is responsible for presenting tumor antigens to T cell receptors (Wang et al. (2021)). As such mutations

of B2M facilitate a tumor's escape from immune surveillance. Mutations of CREBBP are known to reduce transcription of MHC-II, which thereby decreases antigen presentation too (Green et al. (2015)). Therefore both events in this cluster allow for immune evasion by comparable means and could potentially be considered interchangeable in cancer progression. This cluster was also close to the mutation events PTPRS (M), NOTCH3 (M) and KMT2D (M), which could hint at a more general relationship between these events.

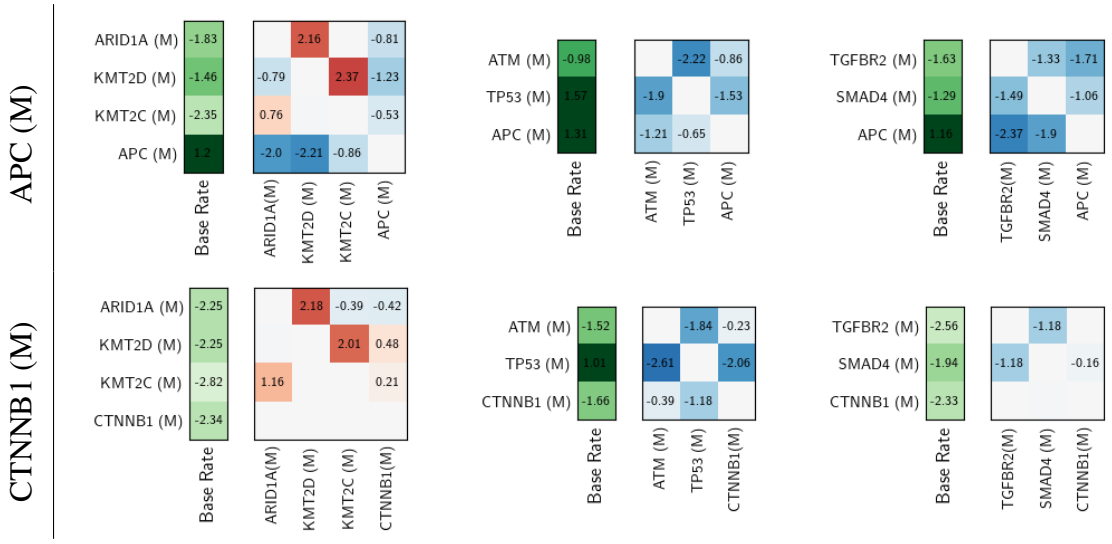
FBXW7 (M) / SOX9 (M):

It has previously been discovered that in the context of DNA damage response FBW7 (also known as FBXW7) directly regulates SOX9 depletion, with GSK3$\beta$ also contributing to this process (Hong et al. (2016)).
This constitutes a strong argument for mutations of FBXW7 and SOX9 indeed being clustered together, because of their biological relation. Additional findings from a study on medulloblastoma indicate that the PI3K/AKT/mTOR signaling cascade is responsible for controlling the aforementioned GSK3-SOX9-FBW7 regulatory axis, which might explain why PIK3CA (M) is also very close to this cluster (Suryo Rahmanto et al. (2016)).

Expected groups not identified by the distance measure:

We also want to mention a negative example of a group of events that we expected to be clustered together based on previous findings that was however not identified as interchangeable by our distance measure.

APC (M) and CTNNB1 (M) are mutations on the genes encoding APC and $\beta$-Catenin, which are both crucial parts of the Wnt/$\beta$-signaling pathway (Nadin et al. (2025)). These mutations have been shown to share similar consequences like the activation of $\beta$-Catenin-Tcf-signaling (Morin et al. (1997)). The large distance between them in our results (and between their $\theta_{T,e}$'s) could however be interpreted as the distance measure $\overline{dist}_{\mathscr{D},\mathbb{T}_{COAD}}$ asserting that other consequences of these mutations are more relevant to the progression of colorectal cancer.



# 5 Conclusion

In this thesis we introduced a new distance measure with the aim of finding clusters of interchangeable cancer progression events based on cross-sectional data of cancer patients' accumulated events. In chapter 2 we constructed the distance measure step-by-step and explained obstacles that we faced - like the difficulties in comparing small MHNs in a meaningful way. Our mathematically concise definition in section 2.5 is meant to underline that - apart from the MHN optimizations - we only make use of very simple tools and methods to achieve our results. We evaluated the distance measure's cluster finding capabilities and justified some non-obvious decisions in the subsequent chapter 3 using a test case with known clustering, which we tried to reproduce under increasingly difficult conditions (more noise and perturbation). The evaluation was however limited to the distance measure with only a single test event set. Expanding it to the combined distance measure with multiple test event sets (as in section 2.6) could prove useful in determining whether the presented method of combining distance measures is actually the best way to recover a given clustering. Additionally this could yield some insight into how the test event sets should best be chosen. In its current form the simulated evaluation showed that our distance measure could successfully retrieve a ground truth clustering of 14 events spread across 5 clusters for sample sizes $\geq 7500$ and perturbation levels $\leq 0.2$. For smaller datasets and higher perturbation levels retrieval was only partially successful including some cases of extreme noise and perturbation, where the predicted clustering hardly outperformed random clusterings.

For the application on selected datasets of cancer patients in chapter 4 this unfortunately meant that the results would only be of limited interpretability as the datasets only contained 2585 and 1697 entries. However there were enough cases of predicted clusters of cancer progression events, which aligned with previous findings on close relationships between these events, to confidently state that our distance measure does have some capabilities in identifying biological relations in cancer progression even for smaller datasets. We argued that some interchangeabilities of cancer progression events may also be explained by their affected proteins taking on equivalent roles in (healthy) cells, which could expand our distance measure's applicability to the study of the more general field of cell signaling. A convincing cluster of such equivalent proteins was already identified with ARID1A and ARID2 in the lung cancer dataset. An example for a cluster of proteins that are known to directly bind to each other was found with FBW7 and SOX9 in the colorectal cancer dataset. What became very apparent during the analysis of this second dataset was that correlated events were also highly likely to get clustered together by our distance measure. Clusters of strongly anticorrelated events were unfortunately not found. This could have hinted at mutually exclusive events with interchangeable roles in cancer progression. Clusters of uncorrelated or weakly correlated events (that additionally had a biological justification) were however present in our results and show that we do not purely rely on direct correlations between events for our findings. Within these clusters the events most

often had similarities in their correlation coefficients with third events. This makes it seem possible that one could have produced similar clusterings by formulating a distance measure that compares two events solely by their correlation coefficients with selected other events.

Some more investigations into the reasonability of the biological statements derived from our identified clusters in chapter 4 would be helpful to judge our presented method more accurately. In general, the study of interchangeability of cancer progression events promises to yield some interesting findings. These may even include (new) information about related pathways and interacting proteins as well as potentially help at understanding which consequences of a genomic event are its most relevant to cancer progression. In this thesis we explored the relevance of these interchangeabilities and presented, evaluated and applied a new statistical method that tries to measure them using MHNs.

# References


Allen, T. D., C. Q. Zhu, K. D. Jones, N. Yanagawa, M.-S. Tsao, and J. M. Bishop (2011). "Interaction between MYC and MCL1 in the genesis and outcome of non–small-cell lung cancer". In: *Cancer research* 71(6), pp. 2212–2221.

Bronder, D., A. Tighe, D. Wangsa, et al. (2021). "TP53 loss initiates chromosomal instability in fallopian tube epithelial cells". In: *Disease models & mechanisms* 14(11), p. dmm049001.

Ciriello, G., M. L. Miller, B. A. Aksoy, Y. Senbabaoglu, N. Schultz, and C. Sander (2013). "Emerging landscape of oncogenic signatures across human cancers". In: *Nature genetics* 45(10), pp. 1127–1133.

Cremona, C. and A. Behrens (2014). "ATM signalling and cancer". In: *Oncogene* 33(26), pp. 3351–3360.

Fodde, R., J. Kuipers, C. Rosenberg, et al. (2001). "Mutations in the APC tumour suppressor gene cause chromosomal instability". In: *Nature cell biology* 3(4), pp. 433–438.

Froimchuk, E., Y. Jang, and K. Ge (2017). "Histone H3 lysine 4 methyltransferase KMT2D". In: *Gene* 627, pp. 337–342.

Green, M. R., S. Kihira, C. L. Liu, et al. (2015). "Mutations in early follicular lymphoma progenitors are associated with suppressed antigen presentation". In: *Proceedings of the National Academy of Sciences* 112(10), E1116–E1125.

Hancock, J. T. (2017). *Cell signalling*. Oxford University Press.

Hodges, C., J. G. Kirkland, and G. R. Crabtree (2016). "The many roles of BAF (mSWI/SNF) and PBAF complexes in cancer". In: *Cold Spring Harbor perspectives in medicine* 6(8), a026930.

Hong, X., W. Liu, R. Song, et al. (2016). "SOX9 is targeted for proteasomal degradation by the E3 ligase FBW7 in response to DNA damage". In: *Nucleic acids research* 44(18), pp. 8855–8869.

Hubert, L. and P. Arabie (1985). "Comparing partitions". In: *Journal of classification* 2(1), pp. 193–218.

Ji, H., J.-H. Lee, Y. Wang, et al. (2016). “EGFR phosphorylates FAM129B to promote Ras activation”. In: *Proceedings of the National Academy of Sciences* 113(3), pp. 644–649.

Jung, J. H., H. Lee, B. Cao, P. Liao, S. X. Zeng, and H. Lu (2020). “RNA-binding motif protein 10 induces apoptosis and suppresses proliferation by activating p53”. In: *Oncogene* 39(5), pp. 1031–1040.

Karakostis, K., L. Malbert-Colas, A. Thermou, B. Vojtesek, and R. Fåhraeus (2024). “The DNA damage sensor ATM kinase interacts with the p53 mRNA and guides the DNA damage response pathway”. In: *Molecular cancer* 23(1), p. 21.

Molina, J. R. and A. A. Adjei (2006). “The Ras/Raf/MAPK Pathway”. In: *Journal of Thoracic Oncology* 1(1), pp. 7–9.

Morin, P. J., A. B. Sparks, V. Korinek, N. Barker, H. Clevers, B. Vogelstein, and K. W. Kinzler (1997). “Activation of $\beta$-catenin-Tcf signaling in colon cancer by mutations in $\beta$-catenin or APC”. In: *Science* 275(5307), pp. 1787–1790.

Morris, L. G., A. M. Kaufman, Y. Gong, et al. (2013). “Recurrent somatic mutation of FAT1 in multiple human cancers leads to aberrant Wnt activation”. In: *Nature genetics* 45(3), pp. 253–261.

Nadin, S. B., F. D. Cuello-Carrión, N. Cayado-Gutiérrez, and M. A. Fanelli (2025). “Overview of Wnt/$\beta$-Catenin Pathway and DNA Damage/Repair in Cancer”. In: *Biology* 14(2), p. 185.

Nowell, P. C. (1976). “The Clonal Evolution of Tumor Cell Populations: Acquired genetic lability permits stepwise selection of variant sublines and underlies tumor progression.” In: *Science* 194(4260), pp. 23–28.

Oikonomou, E., E. Koustas, M. Goulielmaki, and A. Pintzas (2014). “BRAF vs RAS oncogenes: are mutations of the same pathway equal? Differential signalling and therapeutic implications”. In: *Oncotarget* 5(23), p. 11752.

Polager, S. and D. Ginsberg (2009). “p53 and E2f: partners in life and death”. In: *Nature Reviews Cancer* 9(10), pp. 738–748.

Randles, B. M., I. V. Pasquetto, M. S. Golshan, and C. L. Borgman (2017). “Using the Jupyter notebook as a tool for open science: An empirical study”. In: *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, pp. 1–2.

Rivas, S., A. Marín, S. Samtani, E. González-Feliú, and R. Armisén (2022). "MET signaling pathways, resistance mechanisms, and opportunities for target therapies". In: *International Journal of Molecular Sciences* 23(22), p. 13898.

Sansom, O. J., K. R. Reed, A. J. Hayes, et al. (2004). "Loss of Apc in vivo immediately perturbs Wnt signaling, differentiation, and migration". In: *Genes & development* 18(12), pp. 1385–1390.

Schill, R., S. Solbrig, T. Wettig, and R. Spang (2019). "Modelling cancer progression using mutual hazard networks". In: *Bioinformatics* 36(1), pp. 241–249.

Sinclair, A. (2002). "Genetics 101: detecting mutations in human genes". In: *Cmaj* 167(3), pp. 275–279.

Suryo Rahmanto, A., V. Savov, A. Brunner, et al. (2016). "FBW7 suppression leads to SOX9 stabilization and increased malignancy in medulloblastoma". In: *The EMBO journal* 35(20), pp. 2192–2212.

Tanneberger, K., A. S. Pfister, V. Kriz, V. Bryja, A. Schambony, and J. Behrens (2011). "Structural and functional characterization of the Wnt inhibitor APC membrane recruitment 1 (Amer1)". In: *Journal of Biological Chemistry* 286(22), pp. 19204–19214.

Vander Ark, A., J. Cao, and X. Li (2018). "TGF-$\beta$ receptors: In and beyond TGF-$\beta$ signaling". In: *Cellular signalling* 52, pp. 112–120.

Vocht, S. (2022). "Identifiability of Mutual Hazard Networks". (unpublished bachelor's thesis).

Wang, H., B. Liu, and J. Wei (2021). "Beta2-microglobulin (B2M) in cancer immunotherapies: biological function, resistance and remedy". In: *Cancer letters* 517, pp. 96–104.

Wohlhieter, C. A., A. L. Richards, F. Uddin, et al. (2020). "Concurrent mutations in STK11 and KEAP1 promote ferroptosis protection and SCD1 dependence in lung cancer". In: *Cell reports* 33(9).

Wu, R.-C., T.-L. Wang, and I.-M. Shih (2014). "The emerging roles of ARID1A in tumor suppression". In: *Cancer biology & therapy* 15(6), pp. 655–664.

Zhao, M., L. Mishra, and C.-X. Deng (2018). "The role of TGF-$\beta$/SMAD4 signaling in cancer". In: *International journal of biological sciences* 14(2), p. 111.

# Appendix

## A Implementation remarks

This section discusses some more details about the underlying implementations of this thesis and how they can best be accessed and used.

### A.1 General overview

All plots, codes and implementations that were used in this thesis are publicly available at `https://github.com/MichaelBonart/CancerProgressionEvent-Clustering` and `https://github.com/MichaelBonart/BachelorThesis`, where the latter is meant to act as a permanent reflection of the implementation at the time of submission of this bachelor's thesis.

The source code can be found in the directory '**src/**', which contains three subdirectories that correspond to chapters 2, 3 and 4 of this thesis:

- **definition/**: implementation of the distance measure (chapter 2)
- **evaluation/**: implementation of evaluation process (chapter 3)
- **application/**: application on real data of cancer patients (chapter 4)

The central tool that we developed is the **EventDistanceMeasurer**-class, which is located in the 'definitions/'-directory. When using it its following methods should be called in order:

- **EventDistanceMeasurer**$(T, \Omega)$: upon instantiation specify the test event set $T$ and the desired domain of events $\Omega$ on which distance calculations shall be performed.
- **load_data**$(\mathscr{D})$: specify the dataset $\mathscr{D}$ that the distance measure will be based on.
- **train_All_MHNs**$()$: performs necessary MHN optimizations, stores resulting $\theta_{T,e}$'s
- **compute_distance_matrix**$(dist)$: specify method to compare $\theta_{T,e}$'s (as discussed in section 2.4 we recommend using the symmetrizing L1 distance)
  $\rightarrow$ yields distance matrix

The combination of multiple distance measurers using different test event sets (as described in section 2.6) is implemented in '**combineDistMeasurers**$(array[distMeasurer])$', which is a static method of the EventDistanceMeasurer-class. For an example of its usage see 'src/application/genie16_analysis.ipynb', where it is applied on real data of cancer patients.

## A.2 Automated saving and loading of MHN optimization results

The computationally most expensive part of the distance calculation process are the many MHN optimizations done in the method 'train_All_MHNs()'. To avoid performing identical MHN optimizations multiple times, we implemented an automated "checkpoint" system, which checks if an MHN optimization has previously been computed and if so loads its results from storage instead of recomputing it.

As a single distance measure solely depends on its test event set $T$ and its dataset $\mathscr{D}$ (see section 2.5), we use $(hash(\mathscr{D}), T)$ as a key to store and look up previous computation results. Here *hash* is an arbitrary hashing function that depends on the entirety of $\mathscr{D}$. Although multiple datasets $\mathscr{D} \in \{0,1\}^{n \times E}$ could in theory share the same hash value and thus lookup key too, this has not been a relevant concern for the small number of datasets that we used so far (very small in comparison to $|Im(hash)|$).

These features can easily be utilized by making use of EventDistanceMeasurer's child class **EventDistanceMeasurerCP**. Its functionality is identical to its base class, but it incorporates automatic saving of MHN optimization results and loads them again when needed. Setting **FORCE_EXECUTE_COMPUTATIONS** in the **checkpoints_mbonart**-class to True will disable automated loading and recompute needed results upon execution.

## A.3 Recommended workflow

We mainly used jupyter notebooks to execute code as they allow for easy adaptations to plots, parameters and overall structure, while preserving variables. They are generally regarded as a robust tool to document and share code and computation results (Randles et al. (2017)). In our experience however executing computation heavy code in the graphical user interface of jupyter notebooks was tedious and not practical. We therefore suggest the following workflow, which relies on the "checkpoint" system explained in the previous section A.2 and facilitates working with jupyter notebooks in this context.

- **Preparations**: Setup a server for computing and a local device for viewing/analyzing results. They need to be able to transfer files to each other (e.g. by a shared git repository, the two can also just be the same device). Run '**pip install notebook**' in server's console. (This installs the required command 'jupyter execute'.)
- **Implement** scripts in jupyter notebooks while using **EventDistanceMeasurerCP** to automatically enable the "checkpoint" system for all relevant MHN optimizations.
- **Execute** jupyter notebooks on server using the command '**jupyter execute**' in console. All MHN optimization results will be stored as files.
- **Transfer** generated computation result files to local device (using git). Execute jupyter notebook in its graphical user interface. This will be fast and convenient as all needed MHN optimization results can directly be loaded from storage.

# B Choice of test event sets

In this section we want to give a more detailed justification for our choice of test event sets that we used to analyze the datasets in chapter 4. We aimed to group genomic events by biologically related functions, which we will shortly discuss in the following tables:

## B.1 Dataset on lung cancer patients

The following biologically related test event sets were used in section 4.2:

| Test event set | Biological relation | References |
|---|---|---|
| STK11 (M)<br>KEAP1 (M) | mutations promoting ferroptosis protection | Wohlhieter et al. (2020)) |
| FAT1 (M)<br>APC (M) | mutations promoting Wnt-signaling | Morris et al. (2013)<br>Sansom et al. (2004) |
| TP53 (M)<br>RB1 (M)<br>RB1/13q (Del) | regulation of cell cycle progression | Polager and Ginsberg (2009) |
| | regulation of cell cycle progression (via E2F) | |
| EGFR (M)<br>MET (M)<br>KRAS (M)<br>BRAF (M) | activation of Ras-signaling | Ji et al. (2016)<br>Rivas et al. (2022) |
| | members of Ras/Raf-signaling pathway | Molina and Adjei (2006) |

## B.2 Dataset on colorectal cancer patients

The following biologically related test event sets were used in section 4.3:

| Test event set | Biological relation | References |
|---|---|---|
| KRAS (M)<br>BRAF (M) | members of Ras/Raf-signaling pathway | Molina and Adjei (2006) |
| ATM (M)<br>TP53 (M) | DNA damage sensing and response | Karakostis et al. (2024) |
| TGFBR2 (M)<br>SMAD4 (M) | members of TGF-$\beta$-signaling pathway | Vander Ark et al. (2018)<br>Zhao et al. (2018) |
| FAT1 (M)<br>APC (M)<br>CTNNB1 (M)<br>AMER1 (M) | mutation promoting Wnt-signaling | Morris et al. (2013) |
| | members of Wnt/$\beta$-pathway | Nadin et al. (2025) |
| | regulator of Wnt-signaling | Tanneberger et al. (2011) |
| ARID1A (M)<br>KMT2D (M)<br>KMT2C (M) | remodeling chromatin | Hodges et al. (2016) |
| | directly affecting histones | Froimchuk et al. (2017) |

# Eidesstattliche Erklärung

Ich habe die Arbeit selbstständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und bisher keiner anderen Prüfungsbehörde vorgelegt. Außerdem bestätige ich hiermit, dass die vorgelegten Druckexemplare und die vorgelegte elektronische Version der Arbeit identisch sind, dass ich über wissenschaftlich korrektes Arbeiten und Zitieren aufgeklärt wurde und dass ich von den in § 27 Abs. 5 vorgesehen Rechtsfolgen Kenntnis habe.

Regensburg, den 14.08.2025 _____