

Investigating Common Patterns of Cell Type Interaction between tumor subclones and the
Microenvironment in Colorectal Cancer

Introduction

Concerted global sequencing efforts such as The Cancer Genome Atlas have better elucidated the genomic underpinnings and mechanisms of cancer (Cancer Genome Atlas Research Network [CGARN], et al., 2013), allowing for the rapid development of targeted cancer therapies. However, these novel personalized treatments have still only achieved limited levels of success in clinical settings. Acquired drug resistance in particular poses a difficult challenge to treatment efficacy, with recent research suggesting there to be numerous mechanisms of treatment escape tumors can undergo to gain resistance to once effective therapeutic regimens (Nikolaou et al., 2018). One such mechanism is intratumoral heterogeneity, or the presence of multiple cancerous and noncancerous cell subpopulations within a single tumor, as heterogeneity can provide invasive cancers with “significant adaptability” (Ramon y Cajal et al., 2020).

While combating the effects of heterogeneity has proven challenging, recent studies highlight that heterogeneity may also present a weakness that could be exploited by physicians as a treatment option. These studies focus on a field of research known as “evolutionary game theory,” which considers the interactions between and evolution of cell populations as a game, with subclones pursuing certain strategies that present the best options for survival. Utilizing this framework, Wölfl et al. (2021) explain that physicians may be able to direct tumor cells to evolve in predetermined paths that would be detrimental to overall tumor survival or growth. However, while these methods have been proven to exist theoretically, and have shown promise in tumor models, they remain underdeveloped, as limited knowledge of how distinct cell populations interact *in vivo* is available. This paper aims to further develop this unexplored area by analyzing large genomic datasets in the International Cancer Genome Consortium (ICGC) database to identify common patterns of interactions that often arise in colorectal cancer.

Literature Review

Cancer as a Heterogenous Disease

Intratumoral heterogeneity is defined as variations between spatially separated regions or cells in a cancer tumor. While phenotypic heterogeneity has been noted since the conception of cancer research, Nowell (1976) laid the foundation for the study of genetic heterogeneity within primary malignant tumors. He discovered that while cancer tumors start with a few homogenous progenitor cells, the descendents of those cells mutate and duplicate, leading to the formation of multiple subpopulations, known as subclones, with unique mutations that are not present in the tumor as a whole. This results in the growth of multiple cellular subpopulations in a tumor with varying properties and abilities. However, Marusyk et al. (2012) later added to this paradigm proposed by Nowell by explaining that there are many other genetic and nongenetic sources of intratumoral heterogeneity, such as variations in hormonal distributions, extracellular composition, or epigenetic modifications. The researchers notably found that non-tumor cells such as immune or stromal cells, which form what is known as the tumor microenvironment, can be important in directing tumor progression, defining what is now known as microenvironmental heterogeneity.

Recent studies have also suggested that heterogeneity can often exacerbate patient morbidity and mortality. Mroz & Rocco (2013), for example, measured genetic and microenvironmental heterogeneity in a publicly available dataset of 74 head and neck cancers, finding that both correlated with poor health outcomes. Through a thorough review of current literature, Ramón y Cajal et al. (2020) attempted to provide possible explanations for the worsened clinical outcomes highlighted by Mroz & Rocco. For instance, they explain that many precision therapies may only affect molecular targets present in some tumor subpopulations and not others, rendering most traditional treatments only effective against a minority of tumor cells. In addition, they emphasize that heterogeneity also provides tumors with genetic and phenotypic

variation, promoting greater evolutionary plasticity to develop resistance within the tumor.

Through these properties, the researchers claim, heterogeneous tumors can quickly adapt to and survive treatment.

Subclonal Interactions in Genetically Heterogeneous Tumors

While the aforementioned studies identify the impact of increased cell type diversity on tumor complexity and growth, they largely ignore the interactions present between these cell populations. Taking this into account, Hausser & Alon (2020) provide an alternative explanation to Ramón y Cajal's for the detrimental properties of genetically heterogeneous tumors in specific. In their study, they postulate that tumors must perform a set of clear and defined tasks in order to survive and grow, and that as such, individual cells face tradeoffs between specializing in one of these tasks or generalizing in many. Operating under this framework, they computationally evaluate the gene expression profiles of tumors of 15 cancer types to identify how tumors addressed these tasks. From their analysis, the researchers discovered that most tumors were generalists overall but contained clusters of tumor cells that specialized in certain tasks. To explain these findings, they concluded that these specialized subclones in genetically heterogeneous tumors increased the ability of a tumor to adapt to changing task importance and thus respond to treatment.

This idea of intratumoral specialization is further expanded upon by Janiszewska et al. (2019), which identified a pattern of cooperation between distinct genetic subclones and cells in the microenvironment. In this study, researchers manually curated a set of cell lines with various subclonal patterns and then investigated the metastatic potential of tumor grafts derived from each of these cell lines. They found tumors with separate subclones expressing mutations in *FIGF* and *IL11* genes exhibited the most aggressive characteristics. To explain these findings, they then identified that this enhanced tumor growth could be attributed to the fact that the presence of these subclones elevate neutrophil concentration, harming response to metastatic activity.

From this data, they concluded that treatments that could target FIGF-IL11 interactions or elevate neutrophil activity may be promising.

However, Cleary et al. (2014) suggests that directly targeting subclonal cooperation may actually be ineffective. This study found that breast cancer implanted into mouse models often exhibited a biclonal architecture, composed of one clone of HRAS mutant basal cells and another clone of WNT1 producing luminal cells. Under this tumor structure, the HRAS positive basal cells rely on the luminal cells to secrete wnt to maintain tumor growth. When the luminal cells are removed, the basal cells do indeed regress, but only until they either recruit other wnt producing cells or activate dormant wnt producing pathways to substitute for the luminal cells. This indicates that while removing a cooperative subclone may provide short-term benefits, it is not effective as a long-term treatment.

Evolutionary Game Theory

One alternative to these conventional therapeutic options that has recently shown promise in combating intratumoral heterogeneity is evolutionary game theory. Evolutionary game theory focuses on mathematical derivations of cost and benefit to identify the evolutionary path a tumor may follow, and how physicians can modify these costs and benefits to manipulate tumor evolution. Kaznatcheev et al. (2019) demonstrated the feasibility of this strategy in their study where they used molecular assaying to measure subclonal cooperation in non small cell lung cancer, and then treated the tumor with alectinib. They found that when the tumor was untreated, the subclones followed a leader type strategy, where a drug-resistant subclone protected a leader subclone that was non-resistant but more aggressive. However, under treatment with alectinib, metabolic resources increased in scarcity and so the optimal survival strategy switched to competition, leading to an advantage for a resistant, noncooperative, and less aggressive subclone, stunting tumor expansion. Similar positive evolutionary outcomes were replicated in

several other studies analyzing cooperative relationships in subclonal architectures (Archetti & Pienta, 2018; Dingli et al. 2009).

However, while all these aforementioned works demonstrate the capability of evolutionary game theory for long term cancer treatment and care, the field is still underdeveloped and untested in medical settings. One notable barrier to clinical implementation is the fact that many patterns of subclonal interactions and therapy reactions remain unknown. Additionally, there has been limited research into how the tumor microenvironment affects the strategies of tumor subclones. This lack of information could be in large part attributed to the fact that current literature relies on the brute force testing of cancer cell lines or xenograft models to identify cooperative subclones, which requires large amounts of time and resources. This limits the feasibility of discovering a wide array of applicable therapies, especially since each tumor possesses its own unique subclonal structure. While computational methods may be able to streamline this process, most previous studies have not attempted recognize these interactions computationally, in part due to limitations in the ability of cancer data sequencing techniques to detect tumor subpopulations, and in part due to a lack of effective computational analyses to recognize nonlinear relationships within datasets.

Recently, however, with the advent of single cell genomic sequencing and the rising salience of machine learning techniques in biological applications in conjunction with the increasing availability of computational power, computational analyses of evolutionary systems have become increasingly feasible. This forms a gap in current literature that calls for a broader investigation of commonly occurring patterns of interaction in cancerous tumors using trained machine learning models. Through a computational analysis of the single cell RNAseq data from 120 colorectal cancer tumors, this study attempts to address this limitation by asking what common patterns of relationships develop between cell clusters and the tumor microenvironment

in colorectal cancer in order to help guide future research in evolutionary game theory as a possible treatment.

Methods

A. Overview

In order to detect these cell type interactions, this study relies on machine learning techniques to quantitatively measure nonlinearities in tumor subpopulation frequencies in colorectal adenocarcinoma. Machine learning can “learn” nonlinear and multivariable patterns in data, which could not be explained by a traditional regression or correlation analysis, making it perfect to identify complex game-theoretical interactions between patients. The data to perform this analysis is originally sourced from a single-cell RNA database, which is then preprocessed and clustered to derive tumor subpopulations and frequencies. Then the identity of each subpopulation is classified using functional annotation software and literature review, with unclassified populations identified by overexpressed gene networks. The resulting data is then processed using machine learning techniques to identify intratumor relationships.

B. Data Collection and Preprocessing

This study uses single-cell sequencing data, specifically scRNA-seq, to measure cell type frequency in samples. This approach presents many advantages over bulk tumor sequencing, especially in terms of better defining molecular heterogeneity, as single-cell seq presents a more granular view of tumor cells, while bulk seq only presents the average molecular profile of a sample (Rantalainen, 2018). However, single-cell sequencing remains undeveloped in certain applications. Accuracy in scDNA sequencing has been particularly challenging since there are only 2 copies of a DNA sequence within each cell (Rantalainen, 2018). For this reason, this study opts to use scRNA-seq instead. While RNA sequencing is unable to recognize somatic mutations within a genome, it can be used to identify rare cell populations and transcriptional heterogeneity can determine the functional differences between cells (Haque et al., 2017). This

allows further investigation into the relationships between both cancerous and noncancerous cell types such as immune and stromal cells.

At the same time, a major limitation with using single-cell sequencing methods is the limited quantity of public scRNA data available (Chazarra-Gil et al., 2021). To source a large enough dataset to identify nonlinear relationships, the scRNA-seq data from 120 tumor samples of colorectal adenocarcinoma was downloaded from Gene Expression Omnibus (Pelka et al., 2021). The dataset was formatted in the CellRanger hierarchical data format (Zheng et al., 2017), containing level 2 data (raw read counts, genomic barcodes, sample ids). While this study is still limited by the fact that only 120 tumors are analyzed, a comparatively limited dataset compared to bulk tumor, the vastly improved resolution of single-cell technologies justifies relying on a lower quantity of data.

The dataset was then preprocessed using the python library scanpy (Wolf et al., 2017). Low information data was removed from the dataset by filtering out genes present in less than 500 cells and cells expressing less than 200 genes. To maintain high quality cells, cells with over 15 percent mitochondrial RNA composition and less than total reads were excluded. This left 147,658 remaining cells for downstream analysis. Raw read counts were then normalized using scanpy's `normalize()` function and then log transformed. Cell cycle effects were then regressed using a list of cell cycle genes from Kowalczyk et al. (2015). Afterwards, highly variable genes were chosen by filtering genes based on a minimum dispersion value of 0.25, and mean between 0 and 3 log normal read counts. Then scanpy's `pca` function was used to perform dimensionality reduction.

C. Clustering and Data Analysis

To remove any confounding factors between separate tumor samples such as collection error, scanorama was used for batch integration (Hie et al, 2019). Scanorama, in specific, was chosen over other batch integration methods because of its proficiency in working with large

(100,000+ cell) datasets and its ability to accurately maintain biologically confounding factors (Luecken et al., 2021). Next, cell subpopulations were determined from integrated data. For this process, normalized gene expression data for each cell sample was clustered using the leiden clustering function provided by the scanpy and leidenalg libraries. Leiden uses a K-nearest neighbors algorithm, which calculates the distance between two sets of values, to create an indirect, weighted graph, with each cell as a node and edges placed between cells with highly similar RNA expression values. Then a community detection method is applied to this graph, identifying clusters of highly connected and similar cells. Leiden is particularly effective for our purposes of identifying subclonal relationships since it can easily detect small tumor subpopulations (Traag et al., 2019). The algorithm revealed 7 major clusters at 0.2 resolution, and 92 subclusters at a resolution of 1.5. The biological cell types of these clusters were then manually annotated and identified using a list of highly expressed genes for each cell type provided by Pelka et al. (2021). The 7 major clusters were all labeled, while a few subclusters were classified as the same cell type, leading to 87 total distinct cell subtypes (See Appendix A for full list).

Following cell type annotation, the proportion of the whole tumor each cell subtype composed was organized into a dataset. Then, the scipy library in python was used to calculate pairwise mutual information scores. Mutual information measures the KL divergence, or difference, between two probability functions. Low scores can signal a functional relationship between two variables (Tzannes & Noonan, 1973). As such, mutual information was calculated using the `mutual_info_score` function from the sklearn library and relationships with a mutual information score >0.25 were recognized as significant. Using the resulting list of relationships, cell type modules were defined by a cell type with significant mutual information relationships with at least 2 other cell types, and the cell types that the main cell type was related to (See Appendix B for a complete list).

D. Machine Learning

After establishing modules, machine learning methods were used to confirm the significance of the modules and more specifically identify concrete patterns of interaction between cell types. After testing neural network, random forest, and xgboost machine learning model architectures, xgboost trees were chosen as the optimal method for their high accuracy on the limited set of training data available. The XGBRegressor method was applied by selecting the cell type that was common in all significant relationships in a module as the output, and all the other cell types as predictor features. The data was split randomly into training and test sets with 30% of the data allocated to testing, and then the model was trained with max tree_depth of 3, learning rate of 0.01, and a subsample of 0.9. All models with a root mean squared error (RMSE) of less than 0.5 were considered accurate regressors. Additionally, to mitigate the possibility of overfitting, a scenario where the machine learning model may use non relevant/coincidental factors to make predictions due to the limited training data available, models were also filtered by normalized mean absolute error (NMAE), which measures the error of each data point, divided by the value of each datapoint. This guarantees that larger values have larger errors, which is typical in a scenario without overfitting. These cutoffs were chosen in accordance with Chakraborty et al. (2021) which used similar machine learning techniques on lung cancer RNAseq data.

Once xgboost models were trained, the SHAP algorithm was applied to the models to explain their results and provide insight into how the cell types influenced each other (Lundberg & Lee, 2017). SHAP uses a game theoretic concept of Shapley Coefficients to identify how each variable (clone frequency) affects the predictions of a machine learning model. It does this by measuring the extent to which a model's prediction would change when it is excluded from the dataset and identifying how changing a variable changes the model's prediction (Lundberg & Lee, 2017). For example, if in a certain tumor, one cell type had a value of 1 percent, and the

shapley value for that cell type was -9, that means that the fact that the cell type only had a value of 1 percent reduces the model's prediction by 9 percent. Thus, by comparing shapley values as the proportions of a cell type vary across multiple tumors, it would become possible to identify multivariate, nonlinear relationships within a dataset and so would be appropriate to delineate the asymmetric effects of each cell type. To do so, SHAP values were graphed and manually classified into various categories of intercellular relationships. Only relationships between variables with an absolute mean SHAP value greater than 0.15 were considered in order to validate the accuracy of the relationships. The criteria for categorizing these relationships are explained in Appendix E.

Results

Cell Types

The initial analysis and clustering of the RNAseq data revealed 7 main tumor cell types. Of these clusters, most were related to immune function(B, Mast, Plasma, TNKILC, Myeloid) or other parts of the tumor microenvironment (Stromal), with the epithelial cells likely representing the cancerous/mutated populations.

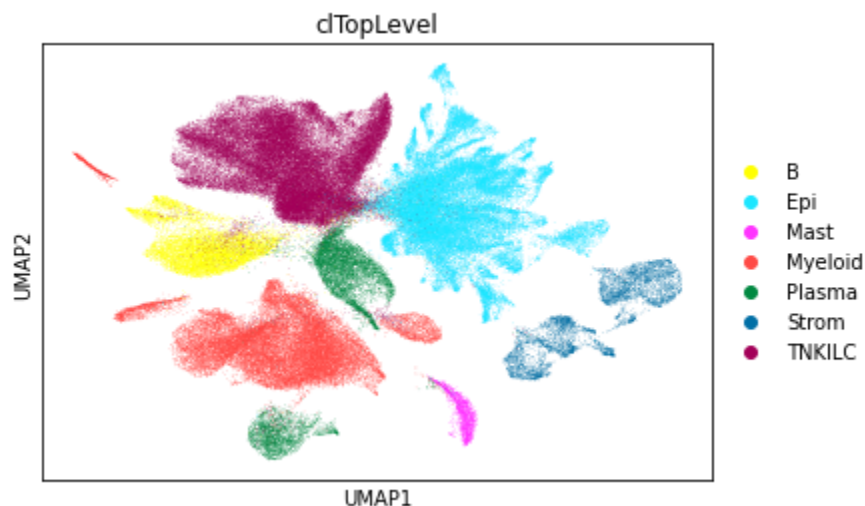


Fig 1. A graphical representation of all single cells and their major cell type categorization. This graph uses Umap, a technique which transforms multi-variable data into a 2 dimensional space for plotting, with highly similar data points plotted closer together.

Of the 87 cell subtypes, however, 33 were stromal subtypes and 26 were TNILKC subtypes, with only 11 epithelial subtypes. Unfortunately, this limited number of epithelial cell subtypes identified limits this study's ability to define many game-theoretically applicable interactions between cancerous populations, but due to the resolution at which the tumor microenvironment was captured in this data, it is feasible to identify a multitude of relationships between various components of the microenvironment and cancerous cells.

Most stromal subtypes were endothelial cells, emphasizing highly heterogeneous blood vessel formations. For the immune cells, it is expected for TNKILC cells to have the most subtypes as the TNKILC cluster represents T cells, Natural Killer cells, and T Natural Killer cells, which are highly diverse categories of immune cells that are very active in tumor development and control (Maccalli et al., 2008).

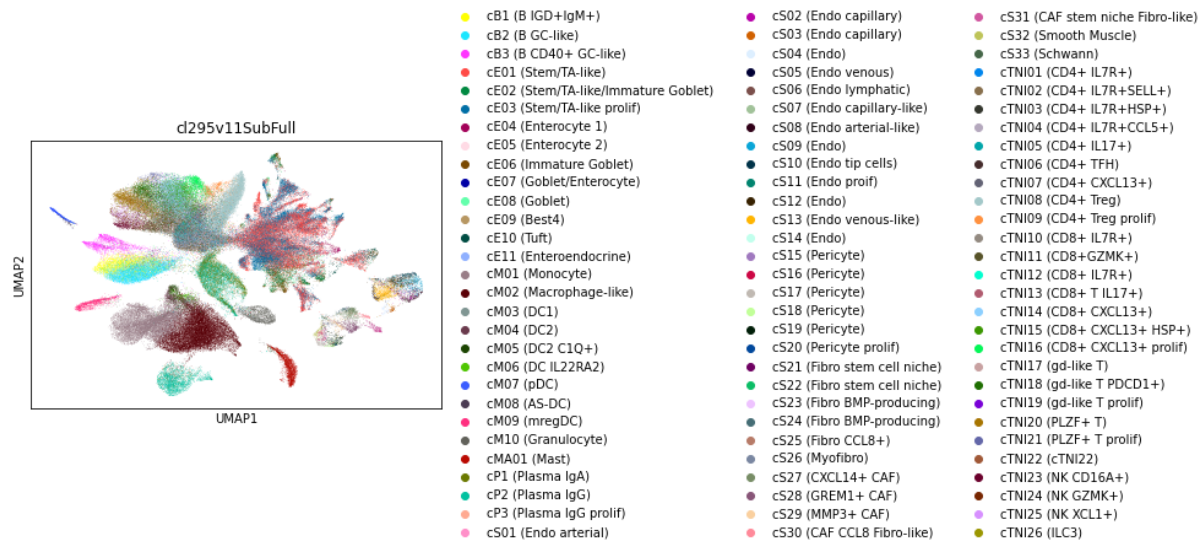


Fig 2. A graphical representation of all single cells and their cell subtype categorization. The subtypes are named using the following conventions: they start with “c” for cell, then the uppercase version of the first letter (first 2 letters in the cast of Mast and TNKILC) of their main cell type, and then an identifying number to distinguish subtypes.

The largest cell subtypes were cE01 (Stem/TA-like) and cE03 (Stem/TA-like prolif) which had mean frequencies of 18.82 and 12.19 percent, respectively. These two groups likely represent cancerous subclones, as stem-like or stem-derivative cells like transit amplifying (TA)

cells tend to drive cancer tumor growth (Barker et al., 2009). Additionally, epithelial cells as a whole represented a large frequency of cells (mean:0689). The next most frequent cell type was TNKILC with a mean of 25.526583 percent. Stromal cells, on the other hand, had a mean frequency of 3.086560 percent.

Machine Learning and SHAP Analysis

Mutual information analysis delineated 46 significant cell type modules. Of the xgboost trees that were constructed from these modules, 29 were “accurate” as per the established requirements ($RMSE < 0.5$, $NMAE < 0.75$). In the case of subtypes cE03 and cE01, larger error margins were allowed ($RMSE < 0.55$, $NMAE < 1.5$) due to the fact that they appear at significantly larger proportions which would affect mean error. In addition to the 46 modules, 11 models were created to relate each epithelial subtype to all other epithelial cell types. The larger error margins used for cE03 and CE01 were also used for these models. Under the assumption that epithelial subtypes represent the mutated or cancerous cell population, these models would measure possible subclonal interactions. Using these models, SHAP analysis demonstrated multiple types of relationships between cell types that could be practically researched for game theoretical treatments.

Generally, all subclonal interactions were classified into 9 relationship types, the criteria for which are explicated in Appendix E: positive, negative, uncorrelated, negative exponential, radical, parabolic, positive bimodal, negative bimodal, and trimodal. Many other models also demonstrated separate regions with varying classifications. For example, some cell types were negatively correlated at low feature values and switched to a positive correlation at a particular inflection point. Unique interaction patterns could also be noticed between epithelial, stromal, and immune cell types.

Examples of Subclonal Interactions

Of the 11 models constructed from the epithelial subtypes to measure subclonal interactions, 6 met the error cutoff necessary to be considered accurate and taken into consideration for SHAP analysis, shown in appendix C.

Consistent with tumor evolutionary theory (Hausser & Alon, 2020), the most closely related cell types, as in the case of the two largest epithelial subpopulations (cE01 and cE03), were linearly and positively related. A similar relationship was also observed between the two enterocyte populations (cE04 and cE05).

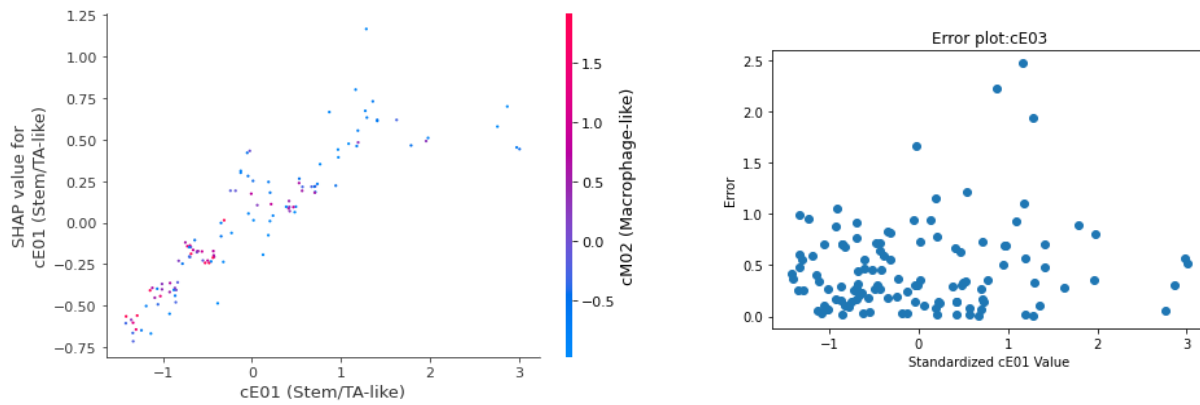


Fig 4. SHAP dependence plot(left) depicting the relationship between cE01 and cE03. Axes are standardized (mean=0, std=1). The two cell types seem to be linearly related except for a group of outliers at high values of cE03. However, the error plot(right) shows that error values remain relatively low at high values of cE03, suggesting this cluster may be phenotypically significant.

However, while the relationships within the enterocyte and stem cell populations were relatively simple, the relationships between them were more complex. For instance, at low values cE04 negatively and linearly affected the model's predictions of cE01, but after the proportion of cE04 in a tumor surpassed $\sim 0.28\%$, the two were positively and linearly related. This is in contrast to cE05 which was positively related to cE01 at low values until valued at $\sim 1.49\%$ at which point the variables become largely unrelated. These asymmetrical responses to cE01, despite cE04 and cE05 being linearly related, not only show that these two subclones vary in phenotypic traits despite their close correlation, but the complexity of their interactions suggest deeper biological meaning.

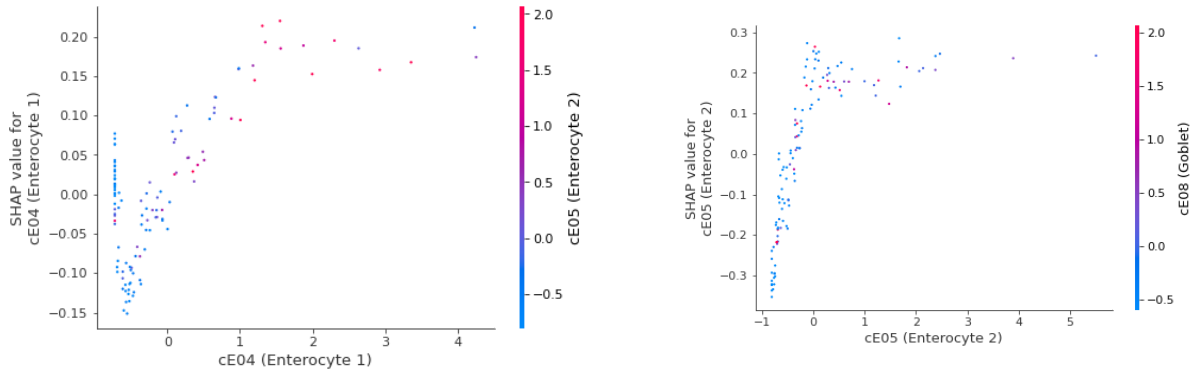


Fig 5. SHAP dependence plots (left) depicting the relationship between cE01 and cE04(left) and cE01 and cE05(right). Axes are standardized (mean=0, std=1).

Similar relationships can be seen between other epithelial cell types such as cE06 and cE08. In the case of cE02(Stem/TA-like/Immature Goblet) and cE01, the shap values depict a negative exponential relationship, with the value of cE01 decreasing at a decreasing rate as the value of cE02 increases. A similar behavior also is observable between cE08(Goblet) and cE05, as the value of cE08 exponentially decreases as cE05 increases. This demonstrates a level of competition between cancerous cell types.

Epithelial cell types are also unique in that they did not demonstrate any bimodal interactions with each other, demonstrating that they likely do not respond to population threshold, which are what likely cause the sudden increase in shap values in bimodal relationships.

Examples of Immune Response

In terms of immune cell types, SHAP analysis revealed what resembles coordinated immune reactions. For example, Modules 0 and 1 (cE03 and cE01) show negative exponential relationships between cE03 and M02(Macrophage-like) and between cE01 and cTNI08(CD4+ Treg). cE05 and cTNI10(CD8+ IL7R+) are also similarly related, revealing a common mode of immune response to epithelial growth.

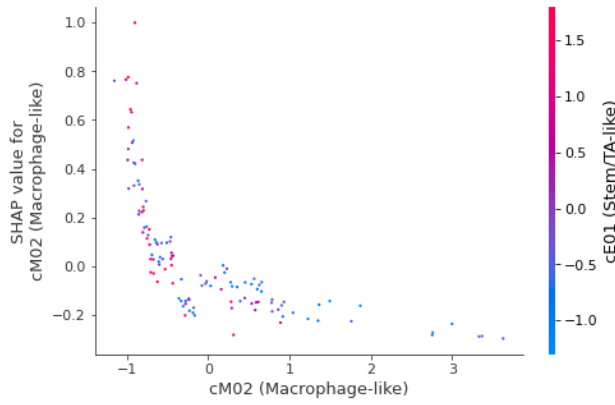


Fig 6. SHAP dependence plot depicting negative exponential relationship between cE01 and cM02. Axes are standardized (mean=0, std=1).

For interactions between immune cell types, a pattern similar to many of the relationships between epithelial cell types emerges where cell types are negatively correlated at low values, while being positively correlated at high values. This is observed between cTNI08 and cTNI04 (CD4+ IL7R+CCL5+), as the two are negatively and linearly related if cTNI04 makes up less than 0.53% of the tumor. Similar interactions can be observed between cM01 and cM02, cM01 and cM03, cB2 and cM09, cTNI10 and cE05, TNI23 and TNI25, and several other cell types.

Interestingly, this type of interaction is the only interaction between immune cells where cell types are negatively related, and only make up 16% of immune-immune relationships compared to 40% of epithelial relationships. All other immune-immune interactions are positive relationships.

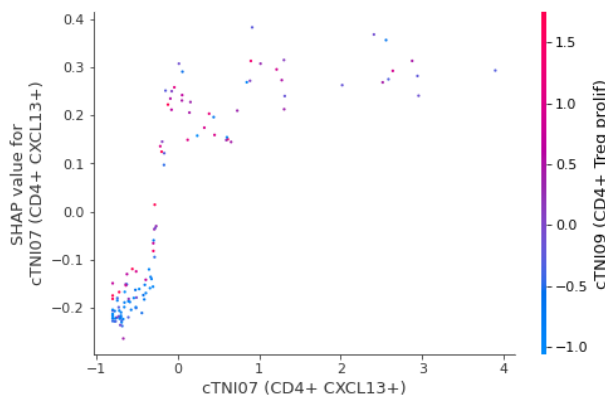


Fig 7. SHAP dependence plot depicting negative exponential relationship between cTNI08 and cTNI07. Axes are standardized (mean=0, std=1).

Other cell types like cTNI08 and cTNI07(CD4+ CXCL13+) demonstrate a bimodality in their relationship, where at low values and high values the two are slightly positively related, but for a certain range of middle values, the two are strongly correlated. This results in two primary states where the value of cTNI08 does not respond to fluctuations in the presence of cTNI07, and a transition phase between the two states, where cTNI08 is incredibly sensitive to changes in cTNI07. cM01 and cM08, and cM09 and cM04 are similarly related.

Examples Stromal Interactions

Many of the interactions between stromal cell types are very similar to those of immune cells. However, in contrast to immune cell types which are mostly positively related, multiple stromal cell types are negatively related to each other. An example of this is the interaction between cS09(endo) and cS17(pericyte), as the two are linearly and negatively related. Other cell types like cS11(endo prolif) and cS09(endo) or cS32(smooth muscle) and cS13(endo venous-like) are positively related at low values and negatively related at high values, in opposition to many of the immune interactions.

Conclusion

Discussion

This study aimed to use machine learning methods to identify common types of interactions between tumor subpopulations that could be used to develop game theory based treatments. Analysis of the generated machine learning models demonstrated 8 such modes of interaction and many other complex interactions that demonstrate the widespread feasibility of game theoretical treatments for colorectal cancer, while also highlighting specific interactions between subclones and the tumor microenvironment that could be further investigated and leveraged.

Most importantly, SHaP analysis of the xgboost trees trained on cell proportion values highlighted the presence of hard inflection points that change the way cell types interact. This

means that in most tumors there are interactions that could be leveraged to manipulate tumor evolution, which stands as the basis for game theoretical treatment. This reveals a new understanding of the application of game theory as many previous studies like Kaznetcheev et al. (2020), Janiszewska (2019), and Wölfl (2021) only highlight the feasibility of game theoretical treatment in certain circumstances, such as when a pair of subclone are present. However, while this study provides evidence that cell type interactions may allow such treatments to be possible in a large variety of tumors, not just in specific circumstances, it does not prove that these interactions can actually be leveraged in a clinical setting by specific treatments, only the possibility that they could.

For instance, many cell types were negatively correlated at low values and positively correlated at high values. This trend was observed between epithelial, immune, and stromal cell populations, demonstrating a nonspecific phenomena. While this pattern could be due to one cell type outcompeting another, no strong conclusions can be made about the underlying biological processes driving this relationship, but either way, this phenomena could be utilized to turn tumor subclones against each other by manipulating relative population sizes, but only if a treatment that could accurately affect subclonal populations could be developed.

This paper also highlights other unique relationships and phenotypic phenomena that emerge due to biological factors that are unique to cell types. For example, the relationship between cS17(Pericyte) and cS09(Endo) is positive at low values of cS17 but later becomes negative, a trend which can be explained by previous literature, as pericytes are known to promote angiogenesis which can harm vessel structure when uncontrolled (Sun et al., 2021). Other findings, such as a cluster of outliers in the relationship between cE01 and cE03 at high values of cE01, require further investigation. Either way, these results highlight avenues for future exploration into tumor environmental behaviors.

Limitations

However, machine learning models still cannot prove causation or concretely define functions to describe the relationship between two variables. Instead they merely highlight a pattern, which must then be confirmed by either in-depth statistical analysis or experimental testing. As both of these methods are unfeasible with the limited scope of data available in this study, the SHAP analysis conducted remains below the burden of proof required to draw strong conclusions. Additionally, the merit of a regressor model is generally determined by previous benchmark models in the current literature, but this is impossible as machine learning has not previously been applied to investigating cell type interactions. While error cutoffs were designed to assure relative accuracy, experimental testing is again required to determine whether the models are accurate enough for practical purposes.

This study is also limited by a lack of available single cell data. A larger sample size of tumor would have greatly improved accuracy. This is further compounded by the fact that single cell RNA sequencing cannot fully capture the genetic differences between tumor subclones, contributing to the fact that many of the epithelial cell type models did not meet the requirements to be considered accurate and could not be considered. Additionally, many of the epithelial cells were grouped into two main subtypes (cE01 and cE03), which have possibly contained smaller subclones that could not be separated at the level of resolution RNAseq provides. Future studies would benefit from combining scRNA and scDNA sequencing rather than just RNAseq to better identify cancerous subclones.

Significance and Future Work

This study is the first to use machine learning to identify interactions between tumor subpopulations, building upon previous work in evolutionary game theory. As such, it also demonstrates the utility of machine learning tools in discovering biological processes within tumor environments and opens the door for future machine learning applications in the field of evolutionary game theory. This study is also one of few to apply game theory to the tumor

microenvironment, and its results add to Kaznatcheev et al. (2019)'s assertion that a heterogeneous tumor microenvironment can also be manipulated to indirectly manipulate subclonal populations, as different sub clones poses different interactions with each part of the tumor microenvironment.

The results also identify areas of key interest for future research into game theoretical treatment of colorectal cancer, such identifying the biological processes that cause certain cell types to switch between positive and negative relationships or experimentally demonstrating the conditions that cause certain immune and stromal populations to rapidly increase in size. Future research is also necessary to confirm the results drawn from these machine learning models are accurate enough to be practically applied to how populations of cell types change when experimentally manipulated.

References

- Archetti, M., & Pienta, K. J. (2018). Cooperation among cancer cells: Applying game theory to cancer. *Nature Reviews Cancer*, 19(2), 110–117.
<https://doi.org/10.1038/s41568-018-0083-7>
- Barker, N., Ridgway, R.A., Es, J.H., Wetering, M.V., Begthel, H., Born, M.V., Danenberg, E., Clarke, A.R., Sansom, O.J., & Clevers, H. (2009). Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature*, 457, 608–611.
- Breiman, L. (2004). Random Forests. *Machine Learning*, 45, 5–32.
- Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., & Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics*, 45(10), 1113–1120. <https://doi.org/10.1038/ng.2764>
- Cao, Y., Wang, X., & Peng, G. (2020). SCSA: A Cell Type Annotation Tool for Single-Cell RNA-seq Data. *Frontiers in Genetics*, 11.
<https://www.frontiersin.org/article/10.3389/fgene.2020.00490>
- Chakraborty, D., Ivan, C., Amero, P., Khan, M., Rodriguez-Aguayo, C., Başağaoğlu, H., & Lopez-Berestein, G. (2021). Explainable artificial intelligence reveals novel insight into tumor microenvironment conditions linked with better prognosis in patients with breast cancer. *Cancers*, 13(14), 3450. <https://doi.org/10.3390/cancers13143450>
- Chazarra-Gil, R., van Dongen, S., Kiselev, V.Y., & Hemberg, M. (2021). Flexible comparison of batch correction methods for single-cell RNA-seq using BatchBench. *Nucleic Acids Research*, 49.
- Cleary, A. S., Leonard, T. L., Gestl, S. A., & Gunther, E. J. (2014). Tumour cell heterogeneity maintained by cooperating subclones in wnt-driven mammary cancers. *Nature*, 508(7494), 113–117. <https://doi.org/10.1038/nature13187>

- Dingli, D., Chalub, F. A., Santos, F. C., Van Segbroeck, S., & Pacheco, J. M. (2009). Cancer phenotype as the outcome of an evolutionary game between normal and malignant cells. *British Journal of Cancer*, *101*(7), 1130–1136. <https://doi.org/10.1038/sj.bjc.6605288>
- Fidler, I. J. (1978). Tumor Heterogeneity and the Biology of Cancer Invasion and Metastasis. *Cancer Research*, *38*(9).
- Guram, K., Kim, S., Wu, V.H., Sanders, P.D., Patel, S.P., Schoenberger, S.P., Cohen, E.E., Chen, S., & Sharabi, A.B. (2019). A Threshold Model for T-Cell Activation in the Era of Checkpoint Blockade Immunotherapy. *Frontiers in Immunology*, *10*.
- Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., ... Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, *184*(13), 3573–3587.e29. <https://doi.org/10.1016/j.cell.2021.04.048>
- Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine*, *9*(1), 75. <https://doi.org/10.1186/s13073-017-0467-4>
- Hausser, J., & Alon, U. (2020). Tumour heterogeneity and the evolutionary trade-offs of cancer. *Nature Reviews Cancer*, *20*(4), 247–257. <https://doi.org/10.1038/s41568-020-0241-6>
- Hie, B.L., Bryson, B.D., & Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature Biotechnology*, *37*, 685–691.
- Hsu, Y., Li, L., & Fuchs, E. (2014). Transit-Amplifying Cells Orchestrate Stem Cell Activity and Tissue Regeneration. *Cell*, *157*, 935–949.
- Janiszewska, M., Tabassum, D. P., Castaño, Z., Cristea, S., Yamamoto, K. N., Kingston, N. L., Murphy, K. C., Shu, S., Harper, N. W., Del Alcazar, C. G., Alečković, M., Ekram, M. B., Cohen, O., Kwak, M., Qin, Y., Laszewski, T., Luoma, A., Marusyk, A., Wucherpennig,

- K. W., ... Polyak, K. (2019). Subclonal cooperation drives metastasis by modulating local and systemic immune microenvironments. *Nature Cell Biology*, 21(7), 879–888. <https://doi.org/10.1038/s41556-019-0346-x>
- Kaznatcheev, A., Peacock, J., Basanta, D., Marusyk, A., & Scott, J. G. (2019). Fibroblasts and alectinib switch the evolutionary games played by non-small cell lung cancer. *Nature Ecology & Evolution*, 3(3), 450–456. <https://doi.org/10.1038/s41559-018-0768-z>
- Kowalczyk, M.S., Tirosh, I., Heckl, D., Rao, T.N., Dixit, A., Haas, B.J., Schneider, R.K., Wagers, A.J., Ebert, B.L., & Regev, A. (2015). Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome research*, 25(12), 1860-72.
- Luecken, M.D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M., Strobl, D., Zappia, L., Dugas, M., Colomé-Tatché, M., & Theis, F.J. (2021). Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19, 41-50.
- Maccalli, C., Scaramuzza, S., & Parmiani, G. (2008). TNK cells (NKG2D+ CD8+ or CD4+ T lymphocytes) in the control of human tumors. *Cancer Immunology, Immunotherapy*, 58, 801-808.
- Marusyk, A., Almendro, V., & Polyak, K. (2012). Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer*, 12(5), 323–334. <https://doi.org/10.1038/nrc3261>
- Mroz, E. A., Tward, A. D., Pickering, C. R., Myers, J. N., Ferris, R. L., & Rocco, J. W. (2013). High intratumor genetic heterogeneity is related to worse outcome in patients with head and neck squamous cell carcinoma. *Cancer*, 119(16), 3034–3042. <https://doi.org/10.1002/cncr.28150>
- Nikolaou, M., Pavlopoulou, A., Georgakilas, A. G., & Kyrodimos, E. (2018). The challenge of drug resistance in cancer treatment: a current overview. *Clinical & Experimental Metastasis*, 35(4), 309–318. <https://doi.org/10.1007/s10585-018-9903-0>

Nowell, P. C. (1976). The clonal evolution of Tumor Cell Populations. *Science*, 194(4260), 23–28. <https://doi.org/10.1126/science.959840>

Ramón y Cajal, S., Sesé, M., Capdevila, C., Aasen, T., de Mattos-Arruda, L., Diaz-Cano, S. J., Hernández-Losa, J., & Castellví, J. (2020). Clinical implications of intratumor heterogeneity: challenges and opportunities. *Journal of Molecular Medicine (Berlin, Germany)*, 98(2), 161. <https://doi.org/10.1007/S00109-020-01874-2>

Sun, R., Kong, X., Qiu, X., Huang, C., & Wong, P.-P. (2021). The emerging roles of pericytes in modulating tumor microenvironment. *Frontiers in Cell and Developmental Biology*, 9. <https://doi.org/10.3389/fcell.2021.676342>

Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Ou Yang, T.-H., Porta-Pardo, E., Gao, G. F., Plaisier, C. L., Eddy, J. A., Ziv, E., Culhane, A. C., Paull, E. O., Sivakumar, I. K. A., Gentles, A. J., Malhotra, R., Farshidfar, F., Colaprico, A., Parker, J. S., ... Mariamidze, A. (2018). The immune landscape of cancer. *Immunity*, 48(4). <https://doi.org/10.1016/j.immuni.2018.03.023>

Wölfl, B., te Rietmole, H., Salvioli, M. et al. The Contribution of Evolutionary Game Theory to Understanding and Treating Cancer. *Dyn Games Appl* (2021). <https://doi.org/10.1007/s13235-021-00397-w>

Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., Gregory, M.T., Shuga, J., Montesclaros, L., Underwood, J.G., Masquelier, D.A., Nishimura, S.Y., Schnall-Levin, M., Wyatt, P.W., Hindson, C.M., ... Bielas, J.H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8.

Appendices

Appendix A. List of Cell Types, their mean frequency, and their standard deviation

	Mean	STD
cE03 (Stem/TA-like prolif)	12.18860367	8.951977817
cE01 (Stem/TA-like)	18.81504212	13.1423138
cE04 (Enterocyte 1)	0.920284321	1.276447939
cM02 (Macrophage-like)	8.648520478	7.536195092
cE05 (Enterocyte 2)	1.311227511	1.637035962
cTNI08 (CD4+ Treg)	4.093503021	2.422076544
cTNI04 (CD4+ IL7R+CCL5+)	1.821368449	1.824029611
cTNI05 (CD4+ IL17+)	1.218758115	1.65535273
cP2 (Plasma IgG)	3.339759579	4.684860407
cS27 (CXCL14+ CAF)	0.2098512427	0.3780207524
cB2 (B GC-like)	3.044096023	3.316405613
cM01 (Monocyte)	7.550839848	7.668361088
cTNI20 (PLZF+ T)	1.415943814	2.32175408
cTNI09 (CD4+ Treg prolif)	0.79199979	0.646407006
cS12 (Endo)	0.2444700515	0.3516174462
cS13 (Endo venous-like)	0.4368292566	0.6027321525
cTNI13 (CD8+ T IL17+)	0.6432036177	1.524502886
cM05 (DC2 C1Q+)	0.3710989987	0.4093536344
cMA01 (Mast)	1.198158686	1.338459369
cTNI11 (CD8+GZMK+)	1.564928036	1.892256421
cS08 (Endo arterial-like)	0.1571131624	0.1665992409
cTNI17 (gd-like T)	0.5255554865	0.8361539672
cTNI10 (CD8+ IL7R+)	0.4679587368	0.5972997935
cE08 (Goblet)	0.5583055175	0.9419853215
cTNI23 (NK CD16A+)	0.5377638157	0.6031265502
cM04 (DC2)	0.367784087	0.352515377
cE06 (Immature Goblet)	2.125800245	2.854779119
cP1 (Plasma IgA)	1.111365327	2.166783051
cM09 (mregDC)	0.7753128861	0.8279648148
cE10 (Tuft)	0.5622674584	1.670512526
cTNI02 (CD4+ IL7R+SELL+)	0.8653587279	1.322542215
cS29 (MMP3+ CAF)	0.1318073756	0.1992953709

cB1 (B IGD+IgM+)	1.366071389	1.848453669
cS09 (Endo)	0.2468215879	0.2853997663
cTNI16 (CD8+ CXCL13+ prolif)	0.9190679228	1.4111553
cE02 (Stem/TA-like/Immature Goblet)	3.761157379	5.544535285
cM06 (DC IL22RA2)	0.1747877437	0.2261716256
cTNI24 (NK GZMK+)	0.4357936069	0.7528278461
cTNI25 (NK XCL1+)	0.2844838401	0.3715010934
cTNI03 (CD4+ IL7R+HSP+)	0.631322801	1.499831823
cS18 (Pericyte)	0.07494904346	0.1281733753
cTNI01 (CD4+ IL7R+)	1.014500894	1.797241572
cP3 (Plasma IgG prolif)	0.2502939578	0.3624267261
cE11 (Enteroendocrine)	0.1472145025	0.3459482458
cS11 (Endo prolif)	0.07205272376	0.1014131773
cTNI07 (CD4+ CXCL13+)	0.9130700882	1.149598692
cTNI14 (CD8+ CXCL13+)	3.014066963	4.957188727
cTNI26 (ILC3)	0.1349340648	0.1658561494
cB3 (B CD40+ GC-like)	1.114118481	2.936463826
cTNI18 (gd-like T PDCD1+)	1.221084024	2.399840977
cS15 (Pericyte)	0.0485420298	0.1142250919
cS02 (Endo capillary)	0.07887698017	0.1506506173
cS17 (Pericyte)	0.130376729	0.1639967177
cM10 (Granulocyte)	0.9010179044	2.295090752
cS04 (Endo)	0.08050622756	0.3212524165
cS19 (Pericyte)	0.1432329758	0.271502126
cTNI06 (CD4+ TFH)	0.3798418178	0.7954851301
cM03 (DC1)	0.1059309169	0.111115095
cS25 (Fibro CCL8+)	0.01454716255	0.03972829286
cS28 (GREM1+ CAF)	0.2724489365	0.6854278769
cE07 (Goblet/Enterocyte)	0.04898850071	0.114025831
cE09 (Best4)	0.1417980531	0.2152050811
cS10 (Endo tip cells)	0.1565469729	0.214430698
cTNI22 (cTNI22)	0.3258965992	0.4608527925
cS32 (Smooth Muscle)	0.06328705866	0.150385965
cTNI21 (PLZF+ T prolif)	0.2458650769	0.4335869451
cM07 (pDC)	0.436683435	0.6507831376
cTNI15 (CD8+ CXCL13+ HSP+)	1.556682018	5.638296638

cS24 (Fibro BMP-producing)	0.05179783774	0.1258592214
cTNI12 (CD8+ IL7R+)	0.3943961882	1.112774615
cS30 (CAF CCL8 Fibro-like)	0.05515934203	0.2112779546
cS01 (Endo arterial)	0.005914531901	0.02347463143
cM08 (AS-DC)	0.05032823639	0.07557489932
cS20 (Pericyte prolif)	0.01733292663	0.04221084599
cS14 (Endo)	0.03928147081	0.06831231402
cTNI19 (gd-like T prolif)	0.1092351432	0.2048696193
cS06 (Endo lymphatic)	0.02323636078	0.07476240144
cS26 (Myofibro)	0.05390446851	0.1381245976
cS16 (Pericyte)	0.04071539826	0.1555601633
cS31 (CAF stem niche Fibro-like)	0.09227418138	0.5681075355
cS22 (Fibro stem cell niche)	0.003182305855	0.01485825198
cS05 (Endo venous)	0.05227838653	0.1079585278
cS33 (Schwann)	0.01751828726	0.05664533379
cS03 (Endo capillary)	0.04095417752	0.09482495979
cS23 (Fibro BMP-producing)	0.01028959759	0.0277530387
cS21 (Fibro stem cell niche)	0.0008564864475	0.005608517753
cS07 (Endo capillary-like)	0.01960481246	0.1003312233

Appendix B. List of All Cell Type Modules

Prediction	Features
cE03 (Stem/TA-like prolif)	['cE01 (Stem/TA-like)', 'cE04 (Enterocyte 1)', 'cM02 (Macrophage-like)', 'cS09 (Endo)']
cE01 (Stem/TA-like)	['cE03 (Stem/TA-like prolif)', 'cTNI08 (CD4+ Treg)', 'cM01 (Monocyte)', 'cS17 (Pericyte)']
cE04 (Enterocyte 1)	['cE03 (Stem/TA-like prolif)', 'cE05 (Enterocyte 2)', 'cP1 (Plasma IgA)', 'cS02 (Endo capillary)', 'cE09 (Best4)', 'cS24 (Fibro BMP-producing)']
cM02 (Macrophage-like)	['cE03 (Stem/TA-like prolif)', 'cM01 (Monocyte)', 'cM05 (DC2 C1Q+)', 'cMA01 (Mast)', 'cM04 (DC2)', 'cM09 (mregDC)', 'cS17 (Pericyte)']
cS09 (Endo)	['cE03 (Stem/TA-like prolif)', 'cTNI04 (CD4+ IL7R+CCL5+)', 'cS27 (CXCL14+ CAF)', 'cS08 (Endo arterial-like)', 'cS29 (MMP3+ CAF)', 'cS18 (Pericyte)', 'cS11 (Endo prolif)', 'cS17 (Pericyte)', 'cS10 (Endo tip cells)']
cTNI08 (CD4+ Treg)	['cE01 (Stem/TA-like)', 'cTNI04 (CD4+ IL7R+CCL5+)', 'cTNI09 (CD4+ Treg prolif)', 'cTNI23 (NK CD16A+)', 'cM04 (DC2)', 'cM09 (mregDC)', 'cTNI25 (NK XCL1+)', 'cTNI07 (CD4+ CXCL13+)']
cM01 (Monocyte)	['cE01 (Stem/TA-like)', 'cM02 (Macrophage-like)', 'cM04 (DC2)', 'cM09 (mregDC)', 'cS17 (Pericyte)', 'cM03 (DC1)', 'cTNI22 (cTNI22)', 'cM08 (AS-DC)']
cS17 (Pericyte)	['cE01 (Stem/TA-like)', 'cM02 (Macrophage-like)', 'cS27 (CXCL14+ CAF)', 'cM01 (Monocyte)', 'cS13

	('Endo venous-like)', 'cS08 (Endo arterial-like)', 'cS29 (MMP3+ CAF)', 'cS09 (Endo)', 'cS18 (Pericyte)', 'cS10 (Endo tip cells)', 'cS32 (Smooth Muscle)', 'cS03 (Endo capillary)']
cE05 (Enterocyte 2)	['cE04 (Enterocyte 1)', 'cTNI10 (CD8+ IL7R+)']
cP1 (Plasma IgA)	['cE04 (Enterocyte 1)', 'cS02 (Endo capillary)', 'cS24 (Fibro BMP-producing)', 'cS05 (Endo venous)', 'cS03 (Endo capillary)']
cS02 (Endo capillary)	['cE04 (Enterocyte 1)', 'cP2 (Plasma IgG)', 'cTNI17 (gd-like T)', 'cP1 (Plasma IgA)', 'cS18 (Pericyte)', 'cS15 (Pericyte)', 'cE07 (Goblet/Enterocyte)', 'cS05 (Endo venous)', 'cS03 (Endo capillary)', 'cS23 (Fibro BMP-producing)']
cE09 (Best4)	['cE04 (Enterocyte 1)', 'cTNI10 (CD8+ IL7R+)']
cS24 (Fibro BMP-producing)	['cE04 (Enterocyte 1)', 'cP1 (Plasma IgA)', 'cS18 (Pericyte)', 'cS25 (Fibro CCL8+)', 'cS05 (Endo venous)', 'cS03 (Endo capillary)', 'cS23 (Fibro BMP-producing)']
cM05 (DC2 C1Q+)	['cM02 (Macrophage-like)', 'cTNI09 (CD4+ Treg prolif)', 'cM04 (DC2)', 'cM09 (mregDC)', 'cM06 (DC IL22RA2)', 'cS18 (Pericyte)']
cMA01 (Mast)	['cM02 (Macrophage-like)', 'cB2 (B GC-like)', 'cM04 (DC2)']
cM04 (DC2)	['cM02 (Macrophage-like)', 'cTNI08 (CD4+ Treg)', 'cTNI04 (CD4+ IL7R+CCL5+)', 'cM01 (Monocyte)', 'cTNI09 (CD4+ Treg prolif)', 'cM05 (DC2 C1Q+)', 'cMA01 (Mast)', 'cTNI23 (NK CD16A+)', 'cM09 (mregDC)', 'cS18 (Pericyte)', 'cM03 (DC1)', 'cM07 (pDC)', 'cM08 (AS-DC)']
cM09 (mregDC)	['cM02 (Macrophage-like)', 'cTNI08 (CD4+ Treg)', 'cM01 (Monocyte)', 'cM05 (DC2 C1Q+)', 'cM04 (DC2)', 'cM03 (DC1)', 'cM08 (AS-DC)']
cTNI10 (CD8+ IL7R+)	['cE05 (Enterocyte 2)', 'cTNI17 (gd-like T)', 'cE09 (Best4)']
cTNI04 (CD4+ IL7R+CCL5+)	['cTNI08 (CD4+ Treg)', 'cM04 (DC2)', 'cS09 (Endo)']
cTNI09 (CD4+ Treg prolif)	['cTNI08 (CD4+ Treg)', 'cM05 (DC2 C1Q+)', 'cM04 (DC2)', 'cTNI07 (CD4+ CXCL13+)']
cTNI23 (NK CD16A+)	['cTNI08 (CD4+ Treg)', 'cM04 (DC2)', 'cTNI25 (NK XCL1+)', 'cM03 (DC1)', 'cM07 (pDC)']
cTNI25 (NK XCL1+)	['cTNI08 (CD4+ Treg)', 'cTNI23 (NK CD16A+)', 'cTNI24 (NK GZMK+)', 'cTNI07 (CD4+ CXCL13+)', 'cM03 (DC1)', 'cM07 (pDC)']
cTNI07 (CD4+ CXCL13+)	['cTNI08 (CD4+ Treg)', 'cTNI09 (CD4+ Treg prolif)', 'cTNI16 (CD8+ CXCL13+ prolif)', 'cTNI25 (NK XCL1+)', 'cTNI14 (CD8+ CXCL13+)']
cP2 (Plasma IgG)	['cP3 (Plasma IgG prolif)', 'cS02 (Endo capillary)']
cS27 (CXCL14+ CAF)	['cS09 (Endo)', 'cS18 (Pericyte)', 'cS17 (Pericyte)']
cS18 (Pericyte)	['cS27 (CXCL14+ CAF)', 'cM05 (DC2 C1Q+)', 'cS08 (Endo arterial-like)', 'cM04 (DC2)', 'cS09 (Endo)', 'cS02 (Endo capillary)', 'cS17 (Pericyte)', 'cS24 (Fibro BMP-producing)', 'cS05 (Endo venous)', 'cS03 (Endo capillary)']
cB2 (B GC-like)	['cMA01 (Mast)', 'cB1 (B IGD+IgM+)']
cB1 (B IGD+IgM+)	['cB2 (B GC-like)', 'cS05 (Endo venous)']
cM03 (DC1)	['cM01 (Monocyte)', 'cTNI23 (NK CD16A+)', 'cM04 (DC2)', 'cM09 (mregDC)', 'cTNI25 (NK XCL1+)', 'cM07 (pDC)', 'cM08 (AS-DC)']

cTNI22 (cTNI22)	['cM01 (Monocyte)', 'cM08 (AS-DC)']
cM08 (AS-DC)	['cM01 (Monocyte)', 'cM04 (DC2)', 'cM09 (mregDC)', 'cM03 (DC1)', 'cTNI22 (cTNI22)']
cS13 (Endo venous-like)	['cS17 (Pericyte)', 'cS32 (Smooth Muscle)']
cS32 (Smooth Muscle)	['cS13 (Endo venous-like)', 'cS17 (Pericyte)', 'cS10 (Endo tip cells)', 'cS16 (Pericyte)']
cS08 (Endo arterial-like)	['cS29 (MMP3+ CAF)', 'cS09 (Endo)', 'cS18 (Pericyte)', 'cS11 (Endo prolif)', 'cS17 (Pericyte)']
cS29 (MMP3+ CAF)	['cS08 (Endo arterial-like)', 'cS09 (Endo)', 'cS17 (Pericyte)']
cS11 (Endo prolif)	['cS08 (Endo arterial-like)', 'cS09 (Endo)', 'cS10 (Endo tip cells)']
cTNI17 (gd-like T)	['cTNI10 (CD8+ IL7R+)', 'cS02 (Endo capillary)', 'cS03 (Endo capillary)']
cS03 (Endo capillary)	['cTNI17 (gd-like T)', 'cP1 (Plasma IgA)', 'cS18 (Pericyte)', 'cS15 (Pericyte)', 'cS02 (Endo capillary)', 'cS17 (Pericyte)', 'cS25 (Fibro CCL8+)', 'cS24 (Fibro BMP-producing)', 'cS14 (Endo)', 'cS05 (Endo venous)', 'cS23 (Fibro BMP-producing)']
cE06 (Immature Goblet)	['cE08 (Goblet)', 'cE02 (Stem/TA-like/Immature Goblet)']
cM07 (pDC)	['cTNI23 (NK CD16A+)', 'cM04 (DC2)', 'cTNI25 (NK XCL1+)', 'cM03 (DC1)']
cS05 (Endo venous)	['cP1 (Plasma IgA)', 'cB1 (B IGD+IgM+)', 'cS18 (Pericyte)', 'cS02 (Endo capillary)', 'cS24 (Fibro BMP-producing)', 'cS03 (Endo capillary)']
cS10 (Endo tip cells)	['cS09 (Endo)', 'cS11 (Endo prolif)', 'cS17 (Pericyte)', 'cS19 (Pericyte)', 'cS32 (Smooth Muscle)']
cTNI16 (CD8+ CXCL13+ prolif)	['cTNI07 (CD4+ CXCL13+)', 'cTNI14 (CD8+ CXCL13+)']
cTNI14 (CD8+ CXCL13+)	['cTNI16 (CD8+ CXCL13+ prolif)', 'cTNI07 (CD4+ CXCL13+)']
cS15 (Pericyte)	['cS02 (Endo capillary)', 'cS03 (Endo capillary)']
cS23 (Fibro BMP-producing)	['cS02 (Endo capillary)', 'cS25 (Fibro CCL8+)', 'cS24 (Fibro BMP-producing)', 'cS03 (Endo capillary)']
cS25 (Fibro CCL8+)	['cS24 (Fibro BMP-producing)', 'cS03 (Endo capillary)', 'cS23 (Fibro BMP-producing)']

Appendix C. List of Significant Epithelial Modules

Feature	Predictions
cE01	[cE01, cE02, cE03, cE04, cE05, cE06, cE07, cE08, cE09, cE10, cE11]
cE02	[cE01, cE02, cE03, cE04, cE05, cE06, cE07, cE08, cE09, cE10, cE11]
cE03	[cE01, cE02, cE03, cE04, cE05, cE06, cE07, cE08, cE09, cE10, cE11]
cE04	[cE01, cE02, cE03, cE04, cE05, cE06, cE07, cE08, cE09, cE10, cE11]
cE05	[cE01, cE02, cE03, cE04, cE05, cE06, cE07, cE08, cE09, cE10, cE11]
cE06	[cE01, cE02, cE03, cE04, cE05, cE06, cE07, cE08, cE09, cE10, cE11]

cE08	[cE01, cE02, cE03, cE04, cE05, cE06, cE07, cE08, cE09, cE10, cE11]
------	--------------------------------------------------------------------

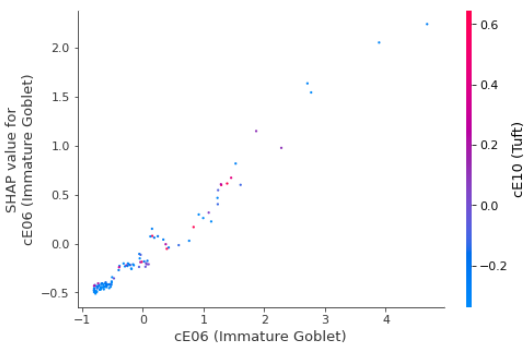
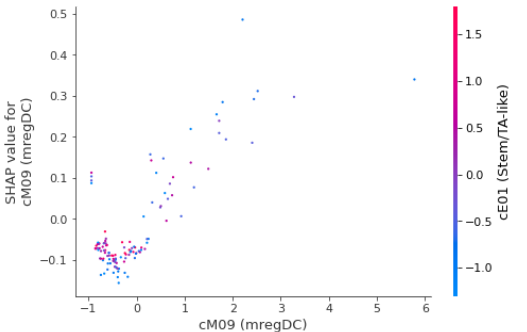
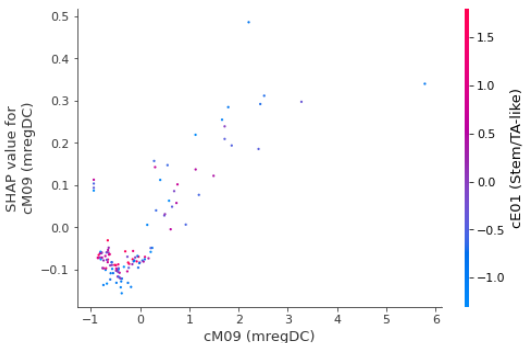
Appendix D. SHAP Analysis Classifications of Significant Models and Relationships with SHAP Mean Absolute Values > 0.15.

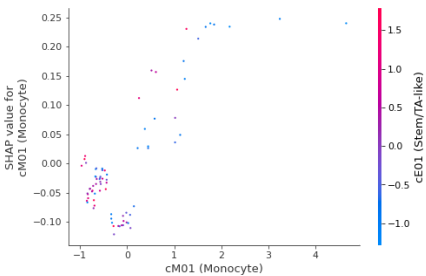
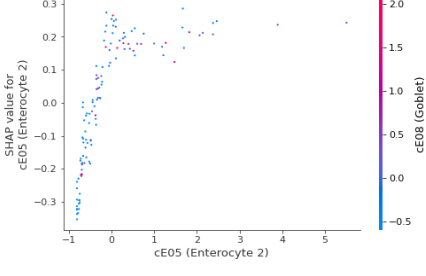
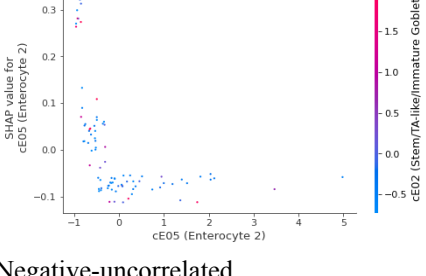
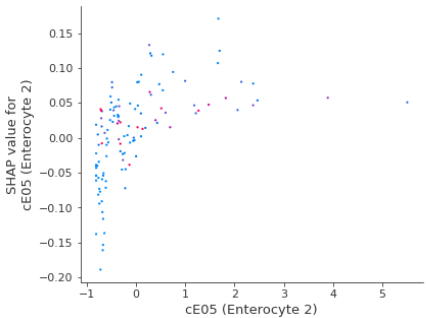
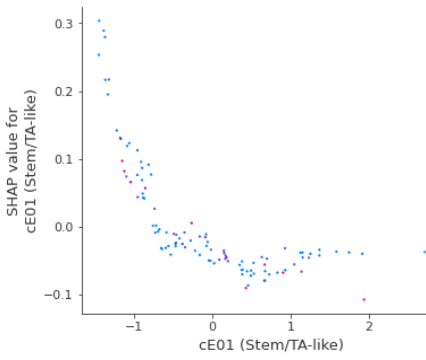
Prediction	SHAP Analysis Features	Classifications
cE03 (Stem/TA-like prolifer)	['cE01 (Stem/TA-like)', 'cE04 (Enterocyte 1)', 'cM02 (Macrophage-like)', 'cS09 (Endo)']	[linear, negative-positive, negative exponential, negative-positive]
cE01 (Stem/TA-like)	['cE03 (Stem/TA-like prolifer)', 'cTNI08 (CD4+ Treg)']	[linear, negative-exponential]
cE04 (Enterocyte 1)	['cE03 (Stem/TA-like prolifer)', 'cE05 (Enterocyte 2)', 'cS02 (Endo capillary)', 'cS24 (Fibro BMP-producing)']	[negative-positive, positive, uncorrelated-positive, radical]
cS09 (Endo)	['cTNI04 (CD4+ IL7R+CCL5+)', 'cS08 (Endo arterial-like)', 'cS18 (Pericyte)', 'cS11 (Endo prolifer)', 'cS10 (Endo tip cells)']	[parabolic-uncorrelated, positive-uncorrelated/bimodal positive-like, positive-uncorrelated/bimodal positive-like, bimodal positive, positive-uncorrelated/bimodal positive-like]
cTNI08 (CD4+ Treg)	['cTNI04 (CD4+ IL7R+CCL5+)', 'cTNI09 (CD4+ Treg prolifer)', 'cTNI23 (NK CD16A+)', 'cM04 (DC2)', 'cTNI07 (CD4+ CXCL13+)']	[negative-positive, vertical asymptote, bimodal positive, radical, bimodal positive]
cM01 (Monocyte)	['cE01 (Stem/TA-like)', 'cM02 (Macrophage-like)', 'cM03 (DC1)', 'cM08 (AS-DC)']	[parabolic, linear, bimodal positive, bimodal positive]
cS17 (Pericyte)	['cS27 (CXCL14+ CAF)', 'cS13 (Endo venous-like)', 'cS09 (Endo)', 'cS18 (Pericyte)', 'cS10 (Endo tip cells)', 'cS32 (Smooth Muscle)']	[linear, bimodal positive, negative-positive, radical]
cM09 (mregDC)	['cM02 (Macrophage-like)', 'cM01 (Monocyte)', 'cM04 (DC2)', 'cM03 (DC1)']	[negative-positive, linear, bimodal positive, linear]
cTNI10 (CD8+ IL7R+)	['cE05 (Enterocyte 2)']	[negative exponential]
cTNI23 (NK CD16A+)	['cTNI08 (CD4+ Treg)', 'cTNI25 (NK XCL1+)']	[trimodal, linear]
cTNI25 (NK XCL1+)	['cTNI23 (NK CD16A+)', 'cTNI24 (NK GZMK+)', 'cTNI07 (CD4+ CXCL13+)']	[negative-positive, negative-positived, uncorrelated-linear]
cP2 (Plasma IgG)	['cP3 (Plasma IgG prolifer)', 'cS02 (Endo capillary)']	[radical, negative-positive-uncorrelated]
cB1 (B IGD+IgM+)	['cB2 (B GC-like)', 'cS05 (Endo venous)']	[linear, parabolic]
cM03 (DC1)	['cM01 (Monocyte)', 'cTNI23 (NK CD16A+)', 'cM04	[negative then positive, bimodal positive,

	(DC2)', 'cM09 (mregDC)', 'cTNI25 (NK XCL1+)', 'cM07 (pDC)', 'cM08 (AS-DC)']	bimodal positive, linear, bimodal positive, radical, bimodal negative]
cTNI22 (cTNI22)	['cM01 (Monocyte)', 'cM08 (AS-DC)']	[linear, radical]
cM08 (AS-DC)	['cM01 (Monocyte)', 'cM04 (DC2)', 'cM09 (mregDC)', 'cTNI22 (cTNI22)']	[bimodal positive, negative-positive-uncorrelated, linear, bimodal positive]
cS13 (Endo venous-like)	['cS17 (Pericyte)', 'cS32 (Smooth Muscle)']	[bimodal positive, linear]
cS32 (Smooth Muscle)	['cS13 (Endo venous-like)', 'cS17 (Pericyte)', 'cS10 (Endo tip cells)', 'cS16 (Pericyte)']	[negative-positive-negative, negative exponential, negative-positive, negative, linear]
cS08 (Endo arterial-like)	['cS29 (MMP3+ CAF)', 'cS09 (Endo)', 'cS18 (Pericyte)', 'cS11 (Endo prolif)', 'cS17 (Pericyte)']	[linear, bimodal positive, linear, negative-positive-uncorrelated, bimodal negative]
cS29 (MMP3+ CAF)	['cS08 (Endo arterial-like)', 'cS09 (Endo)', 'cS17 (Pericyte)']	[parabolic-uncorrelated, negative-positive, trimodal]
cS11 (Endo proilf)	['cS08 (Endo arterial-like)', 'cS09 (Endo)', 'cS10 (Endo tip cells)']	[negative exponential, positive-negative,
cTNI17 (gd-like T)	['cTNI10 (CD8+ IL7R+)', 'cS02 (Endo capillary)', 'cS03 (Endo capillary)']	[negative-positive, trimodal, bimodal positive]
cE06 (Immature Goblet)	['cE08 (Goblet)', 'cE02 (Stem/TA-like/Immature Goblet)']	[linear, radical]
cM07 (pDC)	['cTNI23 (NK CD16A+)', 'cM03 (DC1)']	[bimodal negative, negative exponential]
cS10 (Endo tip cells)	['cS09 (Endo)', 'cS17 (Pericyte)', 'cS19 (Pericyte)', 'cS32 (Smooth Muscle)']	[positive-uncorrelated, trimodal, negative-positive-uncorrelated, bimodal positive]
cTNI16 (CD8+ CXCL13+ prolif)	['cTNI07 (CD4+ CXCL13+)', 'cTNI14 (CD8+ CXCL13+)']	[radical, positive-uncorrelated]
cTNI14 (CD8+ CXCL13+)	['cTNI16 (CD8+ CXCL13+ prolif)', 'cTNI07 (CD4+ CXCL13+)']	[negative-positive, linear]
cS15 (Pericyte)	['cS02 (Endo capillary)', 'cS03 (Endo capillary)']	[bimodal positive, negative-positive-negative]
cS23 (Fibro BMP-producing)	['cS02 (Endo capillary)', 'cS25 (Fibro CCL8+)', 'cS24 (Fibro BMP-producing)', 'cS03 (Endo capillary)']	[trimodal, radical, linear-uncorrelated/bimodal positive-like, positive-negative]
cE01 (epi)	[cE04, cE03, cE05, cE08, cE02]	[negative-positive, linear, positive-uncorrelated, negative-positive-uncorrelated, negative-uncorrelated]
cE03 (epi)	[cE01, cE05, cE06]	[linear-like/trimodal, radical, negative exponential-like]
cE02 (epi)	[cE01, cE06]	[negative exponential, linear]
cE04 (epi)	[cE03]	[uncorrelated-positive]

cE05 (epi)	[cE04, cE07]	[positive-uncorrelated, linear]
cE08 (epi)	[cE05, cE11, cE09]	[negative-uncorrelated, positive-uncorrelated, linear]

Appendix E. Classification Criteria. Graph values are normalized(mean=0, std=1)

Classification feature	Example	Criteria
Linear(Positive)	 <p>Linear</p>	shap values increase at a consistent rate in relation to feature proportion.
Hyphen(-)	 <p>Negative-Positive</p>	A hyphenated classification indicates multiple regions with varying classifications when separated. Each region must contain a Shap values greater than 10% of the total number of Shap values for the model. The order of the words classification represents the order of the regions in the shap dependency plot from low feature values to high feature values
Negative	 <p>Negative-Positive</p>	Shap values decrease at a consistent rate in relation to feature proportion.

<p>Uncorrelated</p>	<div><p>Negative-positive-uncorrelated</p><div><p>Positive-uncorrelated</p><div></div><p>Negative-uncorrelated</p></div></div>	<p>Shap values remain relatively the same even as the feature proportion changes.</p>
<p>Radical</p>	<div></div>	<p>Shap values increase at a decreasing rate as feature values increase. Differs from positive-uncorrelated due to the lack of defined inflection point.</p>
<p>Negative exponential</p>	<div></div>	<p>Shap values decrease at a decreasing rate as feature values increase. Differs from negative-uncorrelated due to the lack of defined inflection point.</p>

parabolic		Shap values decrease then increase as feature values increase. Differs from negative-uncorrelated-positive due to the lack of defined inflection point.
Positive bimodal		Refers to an uncorrelated or lightly correlated region, followed by a highly positive region, followed by another uncorrelated or lightly correlated region. Each region must contain a Shap values greater than 10% of the total number of Shap values for the model.
Negative bimodal		Refers to an uncorrelated or lightly correlated region, followed by a highly negative region, followed by another uncorrelated or lightly correlated region. Each region must contain a Shap values greater than 10% of the total number of Shap values for the model.
Trimodal		Refers to an uncorrelated or lightly correlated region, followed by a highly positive region, followed by another uncorrelated or lightly correlated region, followed by a highly positive region, followed by another uncorrelated or lightly correlated region. Each region must contain a Shap values greater than 10% of the total number of Shap values for the model.