MLPA: A Multi-scale Digital Twin Framework for Personalized Cancer Simulation and Treatment Optimization

James Gu
AlphaMind Club, LLC
Newark, California, USA
jamesguru77@gmail.com

Jake Y. Chen
Department of Biomedical Informatics and Data Science,
Systems Pharmacology AI Research Center
the University of Alabama at Birmingham
Birmingham, Alabama, US
jakechen@uab.edu

ABSTRACT

We introduce the Multi-level Parameterized Automata (MLPA), an innovative digital twin model that revolutionizes personalized cancer growth simulation and treatment optimization. MLPA integrates macroscopic electronic health records and microscopic genomic data, employing stochastic cellular automata to model tumor progression and treatment efficacy dynamically. This multi-scale approach enables MLPA to simulate complex cancer behaviors, including metastasis and pharmacological responses, with remarkable precision. Our validation using bioluminescent imaging from mice demonstrates MLPA's exceptional predictive power, achieving an improvement in accuracy over baseline models for tumor growth prediction. The model accurately captures tumors' characteristic S-shaped growth curve and shows high fidelity in simulating various scenarios, from natural progression to aggressive growth and drug treatment responses. MLPA's ability to simulate drug effects through gene pathway perturbation, validated through equivalence testing, underscores its potential as a powerful tool for precision oncology. The framework offers a robust platform for exploring personalized treatment strategies, potentially transforming patient outcomes by optimizing therapy based on individual biological profiles. We present the theoretical foundation, implementation, and validation of MLPA, highlighting its capacity to advance the field of computational oncology and foster more effective, tailored cancer treatment solutions. As we progress towards precision medicine, MLPA stands at the forefront, offering new possibilities in cancer modeling and treatment optimization. The code and imaging dataset used is available at https://github.com/alphamind-club/MLPA.

CCS CONCEPTS

Computing methodologies → Modeling methodologies • Applied computing → Life and medical sciences → Computational biology

KEYWORDS

Stochastic cellular automaton, cancer growth, gene regulatory pathways, drug response, simulation model

ACM Reference format:

James Gu and Jake Chen. 2024. MLPA: A Multi-scale Digital Twin Framework for Personalized Cancer Growth and Treatment Simulation. In *Proceedings of ACM Shenzhen conference (SHENZHEN'24). ACM, Shenzhen, China, 8 pages.* https://doi.org/10.1145/xxxxx

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SHENZHEN'24, July 2024, SHENZHEN, China
© 2018 Copyright held by the owner/author(s). 978-1-4503-0000-0/18/06...$15.00
https://doi.org/10.1145/xxxxx

1 Introduction

Cancer's inherent complexity and heterogeneity, marked by distinct genetic mutation patterns and diverse clinical presentations [1, 2, 3], often render traditional drug treatment strategies ineffective, leading to suboptimal patient outcomes [4, 5]. The challenge lies in developing personalized treatment approaches that account for the unique characteristics of each patient's cancer [6, 7]. Recent advances in computational modeling have opened new avenues in precision medicine, offering the potential to simulate disease progression and optimize treatment strategies individually.

Recent advances in computational modeling have significantly enhanced our understanding of cancer progression, particularly using stochastic cellular automata. Studies such as Valentim et al. (2022) [8] highlight the efficacy of cellular automata models in exploring tumor growth dynamics under various scenarios. The integration of machine learning with mathematical modeling has emerged as a powerful approach in personalized medicine, particularly in predicting tumor dynamics under conditions of data sparsity and limited mechanistic knowledge. A recent study by Mascheroni et al. (2021) [9] illustrates this approach by proposing a Bayesian method that couples mathematical modeling with statistical learning to enhance predictive accuracy in clinical settings. The study demonstrated that the integration of machine learning with traditional mathematical models significantly improved prediction accuracy, particularly in cases where patient data were sparse or collected at infrequent intervals.

This paper presents a novel multi-tier mechanism-driven digital twin model for cancer that synergizes molecular, cellular, and physiological data across two levels—macroscopic and microscopic. At its core is the Multi-Level Parameterized Automata (MLPA), a novel computational framework that

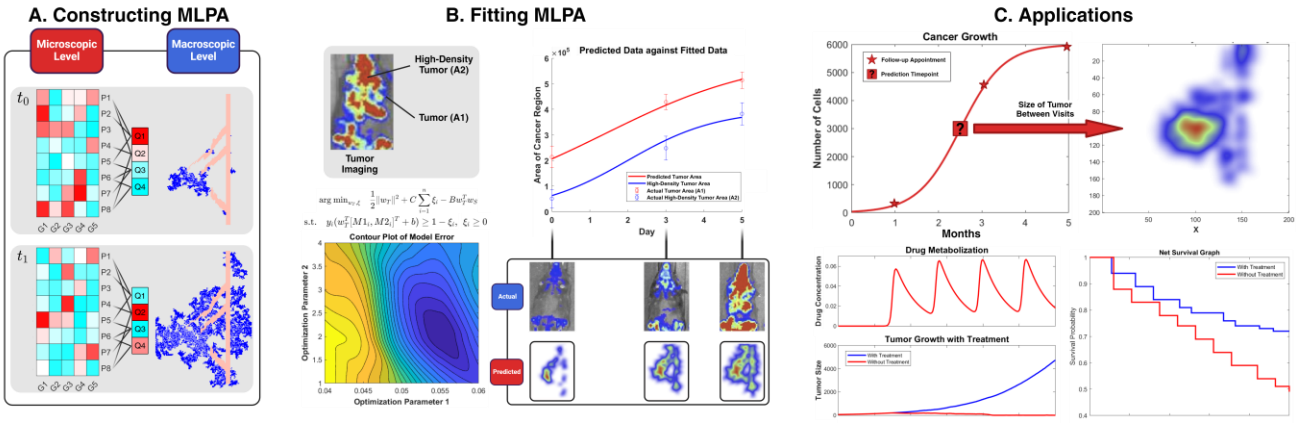

Figure 1: An overview of the MLPA to simulate cancer growth and metastasis. A) The MLPA framework was constructed by integrating data from both microscopic and macroscopic levels. The microscopic level is represented by gene pathways, Q1 to Q4. The macroscopic level represents the spatial distribution and progression of the tumor within the patient's body, highlighting the interactions between cellular automata and angiogenesis over time. As time progresses, changes in gene pathway expression led to dynamic updates in tumor growth and vascular structures. **B)** The MLPA model fitting process involves calibrating the MLPA against bioluminescent imaging data of tumors. Top right: A comparison of predicted versus actual tumor area for high-density and total tumor regions over five days. Bottom left: A contour plot showing the model error landscape for optimization parameters. Bottom right: Visual comparison of actual versus predicted tumor imaging data demonstrates the model's accuracy in replicating tumor growth patterns. Mice images were taken from [19]. **C)** Applications of the MLPA framework can help predict cancer progression and evaluate treatment strategies. Top left: Predicted cancer growth over months, including crucial follow-up appointments and prediction time points. Top right: Spatial heatmap of tumor size between visits. Bottom left: Drug metabolization curves showing drug concentration over time. Bottom middle: Tumor growth comparison with and without treatment. Bottom right: Net survival probability graph illustrating the impact of treatment on patient survival.

facilitates the simulation of cancer proliferation, metastasis, and pharmacological responses [10]. MLPA integrates electronic health records (EHRs) with genomic data, which is increasingly available in precision medicine settings, to develop a detailed, mechanism-based, dynamic model of the tumor microenvironment and the patient's health observations [11, 12, 13].

Our approach is inspired by methodologies like the Swedish Digital Twin Consortium, which constructs patient-specific digital twins incorporating disease-related variables [14]. However, MLPA extends this concept by integrating microscopic-level parameters, providing a more detailed representation of cancer dynamics. By employing stochastic cellular automata, MLPA can model the unpredictable nature of cancer progression while still enabling the identification of actionable therapeutic interventions [15].

The MLPA framework is a virtual testbed for simulating disease progression and evaluating therapeutic responses under various treatment regimens. This in silico approach facilitates an in-depth understanding of treatment outcomes, potential resistance mechanisms, and the optimization of personalized treatment plans. Our modeling strategy parallels computational methods used in other fields, such as those employed to assess insulin resistance using mechanistic mathematical models [16].

In this paper, we present the theoretical foundation of MLPA, its initial prototype, and its validation using bioluminescent imaging data from a mice dataset. We demonstrate MLPA's ability to accurately predict tumor dynamics and its potential for evaluating personalized treatment strategies. By constructing this proof-of-concept personalized cancer digital twin, our work aims to advance the frontier of precision medicine and improve the specificity and efficacy of oncological therapies.

The remainder of this paper is organized as follows: Section 2 details the methods used in constructing and parameterizing MLPA. Section 3 presents the results of our model fitting, validation, and simulations of drug effects. Finally, Sections 4 and 5 discuss the implications of our findings, the current approach's limitations, and future research directions in this field.

2 Materials and Methods

2.1 MLPA Framework Overview

The research methodology employs a dual-level approach to create a personalized cancer digital twin to simulate disease progression and optimize treatment strategies. At the macro level, MLPA focuses on simulating the physical aspects of cancer progression within the patient's body using imaging and EHR data. This includes accurately mapping the spatial location and growth dynamics of tumors. Tumor growth is simulated using stochastic cellular automata, where cells within a defined neighborhood have probabilities of becoming tumor cells based on local conditions and growth potentials. Local factors like nutrient availability and proximity to blood vessels influence growth rates. At the micro level, the model integrates genomic data at the molecular level to simulate critical gene pathways affecting tumor behavior, including growth, mutation, and metastasis. This molecular data enables bi-

directional communication at the macro level, influencing parameters such as tumor growth probabilities and metastasis rates (**Fig 1A**).

Model calibration is achieved by fitting the simulation outputs to real data from bioluminescent imaging captured over multiple time points. Monte Carlo simulations are utilized to explore parameter spaces, identifying optimal values that minimize error between simulated and observed data, enhancing the accuracy of the digital twin (**Fig 1B**). Model validation is done with bioluminescent cancer imaging because it resembles the visual output of MLPA. Bioluminescent imaging is valued for its real-time, non-invasive visualization capabilities in preclinical cancer studies [17]. This technique extracts key metrics like cancer area and high-density tumor regions, essential for model validation. These metrics are crucial for validating our model by directly comparing the continuous tumor growth predictions from our Monte Carlo simulation with the observed data.

This digital twin model serves as a virtual platform for personalized medicine applications, facilitating the evaluation of treatment efficacies and the prediction of individualized patient responses to various therapies. This approach provides valuable insight into the future progression of the disease under various treatment scenarios. It allows for a better understanding of how different therapies may impact tumor growth, patient survival, and drug metabolization over time (**Fig 1C**).

2.2 MLPA Model Construction

The construction of the MLPA model involves three core components: the macroscopic level, the microscopic level, and the AI mapping module (**Fig 2)**. Each of these components plays a crucial role in simulating cancer progression and treatment responses.

At the macroscopic level, the model simulates the physical aspects of cancer within the patient's body, such as tumor growth and spread. It integrates EHR data, including variables like patient age, gender, BMI, and other relevant health metrics. This component is designed to reflect how these macroscopic factors influence cancer progression, providing a broader context for the simulation.

At the microscopic level, the model focuses on simulating molecular and cellular processes, including gene pathway activity that influences tumor behavior, such as growth, mutation, and metastasis. This component leverages genomic data to simulate the microscopic environment of the tumor, providing a detailed view of the internal mechanisms driving cancer progression.

The AI mapping module serves as the bridge between the macroscopic and microscopic levels. In this study, we used ChatGPT to map the relationships between gene pathways and model parameters. According to a recent study exploring the use of ChatGPT in the Reactome curation process, ChatGPT was able to efficiently identify gaps in pathway annotations and enhance the enrichment process, highlighting its potential as an asset in automating and refining complex data-driven tasks [18].

MLPA employs a stochastic cellular automata model to simulate cancer progression. The model, built inside the

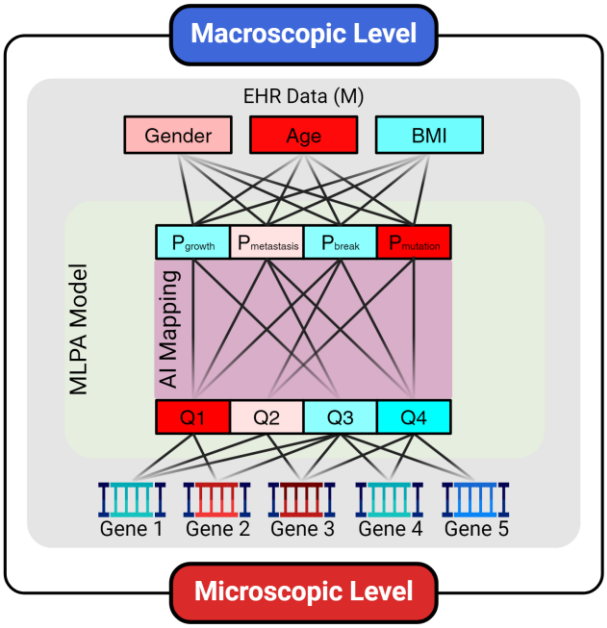

Figure 2: Visualization of key processes in cancer progression modeled with MLPA. The model distinguishes between macroparameters, such as EHR data, and microparameters, such as gene-level data. Macroparameters directly influence the model parameters, while microparameters impact gene pathway expression, with gene pathways denoted as Q1 to Q4. The AI mapping finds relationships between the gene pathways and the model parameters. Increasing intensity of blue indicates higher expression and healthier states, while increasing intensity of red indicates inhibition and less healthy states.

programming language MATLAB, captures the complex dynamics of tumor growth, mutation, and metastasis. The simulation environment is represented by a lattice defined as $L = \{(x,y) \mid 1 \leq x \leq 200, 1 \leq y \leq 200\}$, initialized with a tumor located at the center. Because a two-dimensional lattice is used, a type of search called Moore's neighborhood, defined by $N^M(x_0, y_0) = \{(x,y): |x - x_0| \leq r, |y - y_0| \leq r\}$, can be employed to model tumor growth dynamics. Tumor cells grow based on probabilistic rules within Moore's neighborhood of within a distance $r(u,v) := \max(|u_i - v_i|)$. This model can be generalized to 3D by extending the lattice to include a third dimension, transforming it into a cubic lattice where each cell interacts with its neighbors in all three spatial directions. However, we opt to use a 2D model for several reasons. The primary consideration is computational efficiency; 2D models demand significantly less memory and processing power resources, facilitating faster simulations and enabling multiple iterations to explore different scenarios. This efficiency is crucial for initial studies where the primary goal is to test and refine the fundamental rules and parameters governing tumor dynamics. Additionally, 2D models are simpler to visualize and analyze, providing straightforward insights into the interactions and effects of various factors on tumor growth.

Parameter	Description
P_{growth}	Probability of cancer cell growing each iteration
$P_{metastasis}$	Probability of cancer cells spreading to other parts of the body for each iteration
P_{break}	Probability of blood vessels to undergo EMT and grow towards cancer
$P_{mutation}$	Probability of genetic mutations occurring within cancer cell, changing P_{growth} and $P_{metastasis}$

Adjacent cells within two grid spaces can become tumor cells, with a defined probability P_{growth}. P_{growth} is also influenced by a growth rate map generated using a Gaussian-filtered random map with a standard deviation σ of 1. This map is normalized and polarized to create areas with varying growth potentials, effectively modeling the heterogeneous nature of the tumor microenvironment. Growth rate values range from 0.02 to 1, with higher growth rates assigned to regions with lower polarized map values and vice versa. Cells near blood vessels have their growth probability increased by 50% due to nutrient availability. In the future, we aim to find more imaging datasets to parameterize this variable during the fitting process. An example of how the parameters is used is shown in **Algorithm 1**.

Algorithm 1 Simple cancer algorithm for **MLPA**

for $(x_i, y_i) \in L$ **do**
  if (x_i, y_i) is empty **and** $N^M(x_i, y_i)$ contains cancer
    if $N^M(x_i, y_i)$ contains blood vessel
      $P_{growth} * 1.5$
    else
      P_{growth}
  if (x_i, y_i) is blood vessel
    P_{break}
  if (x_i, y_i) is cancer **and** $N_{cancer} > N_{threshold}$
    $P_{metastasis}$
  if (x_i, y_i) is cancer **and** $P_{mutation}$
    P_{growth}— **and** $P_{metastasis}$++

Blood vessels have a chance P_{break} to grow towards nearby cancer cells. Blood vessels start as rectangular regions from one lattice end. Vessel growth is directed toward Vascular Endothelial Growth Factor (VEGF) signals from tumor cells, simulating angiogenesis. They follow a probabilistic rule, with potential Epithelial-Mesenchymal Transition (EMT) breaks dictated by a decreasing probability P_{break} as breaks accumulate. To enhance nutrient delivery, new vessels grow towards the nearest tumor cell, guided by VEGF signals. Blood vessel offshoots mitigate hypoxia, increasing P_{growth} and $P_{metastasis}$.

Metastasis, defined as $P_{metastasis}$, starts when cancer cells N_{cancer} surpass a defined threshold $N_{threshold}$. This process results in new tumor sites near blood vessels, simulating cancer's vascular spread. Additionally, tumor cells can also mutate with a probability $P_{mutation}$, changing their P_{growth} and $P_{metastasis}$. The model runs for 300 iterations, updating the lattice at each step. Different cancer scenarios are shown in Section 3.1 to show MLPA simulations for normal subtype, aggressive subtype, and drug treated cancer.

To personalize these model parameters to each individual patient, our digital twin integrates macroscopic patient EHR and microscopic patient genomic data to create a highly individualized disease progression model [19]. The macroscopic aspect of our model utilizes EHR data. Blood pressure can be continuously monitored to assess the patient's cardiovascular health, influencing tumor growth, and treatment response [20]. Patient age and gender-specific differences are factored to account for their roles in cancer progression and treatment effectiveness [21, 22, 23]. Other EHR parameters, such as BMI and medical history, are included to comprehensively view the patient's health status. The macroscopic data can directly perturb the stochastic cellular automata parameters such as P_{growth}, P_{break}, $P_{metastasis}$, and $P_{mutation}$, as shown in **Algorithm 2**. However, we acknowledge that the EHR data was not fully tested in the results section due to the use of mice imaging data, which did not come with corresponding EHR data. Moreover, even if such data were available, it would not be directly compatible with human EHR data.

Algorithm 2 Reading macroparameters for **MLPA**

Input: EHR
For attribute $A_i \in$ EHR **do**
  If $A_i > \bar{A}$
    P_{growth} += correlation(A_i, P_{growth})
    $P_{mutation}$ += correlation(A_i, $P_{mutation}$)
    P_{break} += correlation(A_i, P_{break})
    $P_{metastasis}$ += correlation(A_i, $P_{metastasis}$)
  Else if $A_i > \bar{A}$
    P_{growth} -= correlation(A_i, P_{growth})
    $P_{mutation}$ -= correlation(A_i, $P_{mutation}$)
    P_{break} -= correlation(A_i, P_{break})
    $P_{metastasis}$ -= correlation(A_i, $P_{metastasis}$)

On the microscopic level, the genomic pathway data then perturbs the model parameters to tailor the model to individual patients, effectively customizing the simulation based on each patient's specific genomic profile. This perturbation process is guided by the PAGER3 database, which includes pathways scientifically validated to affect metastasis, growth, mutation, and angiogenesis [24, 25, 26]. We extracted pathway data into a CSV file that lists gene pathway names and their interactions to the model parameters. Pathway expression is controlled via a graphical user interface, allowing adjustments based on activation, absence, or inhibition of specific gene pathways. These adjustments directly influence the stochastic cellular automata parameters, as described by their correlation in the CSV file. For instance, an expressed gene pathway associated with increased P_{growth} will adjust the growth rate parameter accordingly, while pathways involved in inhibiting P_{growth} will decrease it, as shown in **Algorithm 3**.

Algorithm 3 Reading microparameters for **MLPA**

Input: Gene Pathway Database
For $Q_i \in$ Gene Pathway Database **do**
  If Q_i = activated
    P_{growth} += correlation(Q_i, P_{growth})
    $P_{mutation}$ += correlation(Q_i, $P_{mutation}$)
    P_{break} += correlation(Q_i, P_{break})
    $P_{metastasis}$ += correlation(Q_i, $P_{metastasis}$)
  Else if P_i = inhibited
    P_{growth} -= correlation(Q_i, P_{growth})
    $P_{mutation}$ -= correlation(Q_i, $P_{mutation}$)
    P_{break} -= correlation(Q_i, P_{break})
    $P_{metastasis}$ -= correlation(Q_i, $P_{metastasis}$)

The AI mapping module was crucial in identifying the relationships between specific gene pathways and the four key model parameters. To achieve this, we utilized ChatGPT, leveraging its capabilities in natural language processing and data interpretation. We prompted ChatGPT using clear and precise details about each pathway to evaluate the potential interactions between each gene pathway and the model parameters while avoiding AI hallucination. Specifically, for each pathway, ChatGPT was asked whether there was a positive reaction, negative reaction, or no reaction concerning the parameters. As the dataset did not include genomic data, we established a standard approach by setting a constant value of a 10% change to the parameters if ChatGPT indicated a certain relationship. This approach ensured that our simulations could still reflect the influence of gene pathways on tumor behavior, even in the absence of direct genomic data.

2.3 Validation Data Acquisition and Preprocessing

The imaging dataset was sourced from the study "Histone H2B ubiquitin ligase RNF20 is required for MLL-rearranged leukemia" (Wang et al., 2013) [27]. This dataset focuses on Mixed-lineage leukemia (MLL) fusions, potent oncogenes that initiate aggressive forms of acute leukemia. MLL-fusion proteins, as aberrant transcriptional regulators, alter gene expression in hematopoietic cells by interacting with the histone H3 lysine 79 (H3K79) methyltransferase DOT1L. Notably, treating leukemia using a DOT1L treatment decreased the spread of the cancer.

The dataset includes bioluminescent imaging data from 16 mice, each recorded at three time points: days 0, 3, and 5. Among these, four mice were treated with a small-molecule inhibitor of DOT1L. We employed a 50/50 training/testing split, using data from 8 untreated mice for model training and data from the remaining 8 mice for testing (4 treated and 4 untreated). In total, 48 images were processed, each tagged with metadata corresponding to the original mouse subject. The images were standardized in size and underwent segmentation based on RGB values to distinguish different regions of the bioluminescent imaging heatmap. For each image, we extracted two critical variables: tumor pixel count (all color values) and high-density tumor pixel count (red color values only).

2.4 MLPA Model Training

From the bioluminescent imaging dataset, we fit third-degree polynomials to the tumor pixel count and high-density tumor pixel count. Harshe et al. (2024) [28] states that three tumor volume measurements are sufficient to estimate model parameters when noise is absent. In our case, we used day 0, day 3, and day 5 imaging data to fit these polynomial curves. These curves provided a mathematical model of the observed growth patterns in the bioluminescent images. Monte Carlo simulations, which are computational techniques that use repeated random sampling to estimate numerical results, were utilized [29]. In the context of MLPA, a Monte Carlo simulation helps explore the range of possible tumor growth scenarios by varying P_{growth} and $P_{metastasis}$ to find which parameter pair most closely represents the cancer growth in the mice dataset. Random values for P_{growth} and $P_{metastasis}$ were chosen from their respective distributions and fed into the stochastic automata model. P_{break} and $P_{mutation}$ were held constant with $P_{break} = 0.001$ and $P_{mutation} = 0.01$. They were held constant as the bioluminescent imaging dataset did not include blood vessel mapping or genomic data to indicate mutations. The pathways from the CSV file, such as mRNA splicing, Processing of Capped Intron Containing Pre mRNA, mRNA Processing, Nonsense Mediated Decay (NMD), Chaperon Mediated Autophagy, Regulation of TP53 Expression and Degradation, and others, were assigned their typical expression in cancerous conditions, with each pathway being assigned a value of 1, -1, or 0. In this context, a value of 1 indicates that the pathway is activated, meaning it is upregulated or more active in the cancerous state. A value of -1 signifies that the pathway is inhibited, reflecting downregulation or reduced activity in the cancerous state. Lastly, a value of 0 indicates that the pathway is not expressed or has a neutral effect in the model. The AI mapping module remained

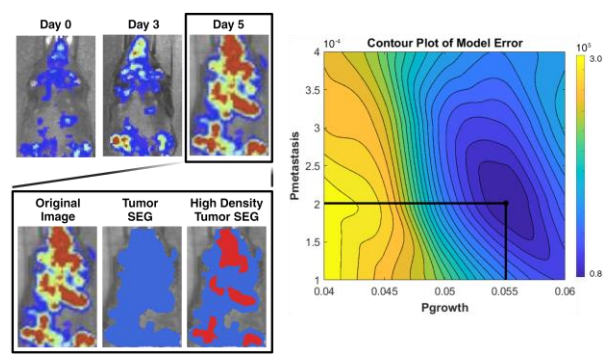

Figure 3: Fitting MLPA to real data involves calibrating the model using bioluminescent imaging data of tumor progression. Tumor growth is shown at days 0, 3, and 5, with an example segmentation of tumor and high-density tumor shown for day 5. A contour plot of model error as a function of growth and metastasis probabilities highlights the optimization process to minimize discrepancies between simulated and observed tumor dynamics. The location of lowest error $\boldsymbol{P_{growth} = 0.055}$ and $\boldsymbol{P_{metastasis} = 0.0002}$ were set to the default model parameters. Mice images were taken from [19].

active throughout this process, ensuring that these values were accurately perturbing the model parameters.

The automata's output was segmented to match the mice images, providing pixel counts for tumor and high-density areas. We compared simulated (derived from MLPA) to observed pixel (derived from mice dataset) counts, calculating residuals for each simulation. Discrepancies were quantified using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Error values were mapped to their respective P_{growth} and $P_{metastasis}$ values (**Fig 3**). The optimal values were determined by identifying the region with the lowest error values, with $P_{growth} = 0.055$ and $P_{metastasis} = 0.0002$. These optimal values became MLPA's default parameters, ensuring accurate tumor growth modeled after the mice dataset. This process underscores the use of a regression model in our approach, where the primary goal is to predict continuous outcomes—in this case, the number of cancer cells over time. By fitting the parameters of MLPA to the observed data, we aim to optimize the model to accurately reflect the tumor growth dynamics observed in the imaging dataset.

3 Results

3.1 MLPA Simulation

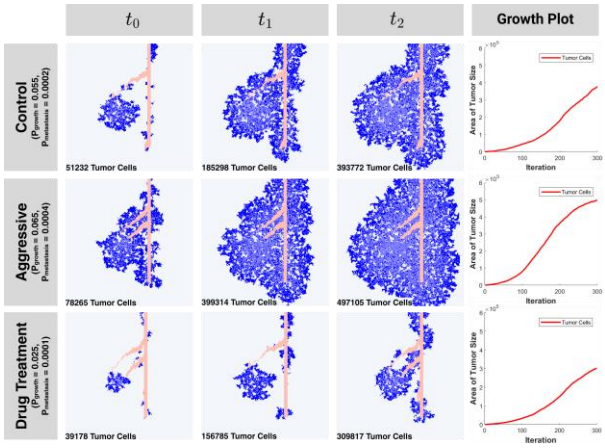

Figure 4: MLPA Simulation of Tumor Growth Under Different Conditions. This figure illustrates the simulated progression of tumor growth under three scenarios: Control, Aggressive, and Drug Treatment. The aggressive scenario has increased $\boldsymbol{P_{growth}}$ and $\boldsymbol{P_{metastasis}}$ due to gene pathways that promote cancer growth that are upregulated. The drug treatment scenario has decreased $\boldsymbol{P_{growth}}$ and $\boldsymbol{P_{metastasis}}$ due to gene pathways that promote cancer growth that are downregulated. Each time point $\boldsymbol{t_i}$ reflects a duration of 100 iterations. The tumor cell count for each iteration has been graphed under the Growth Plot section.

The MLPA model demonstrated its capability to simulate various scenarios of tumor progression, providing insights into different aspects of cancer growth and treatment responses (**Fig 4**).

In the control scenario ($P_{growth} = 0.055$ and $P_{metastasis} = 0.0002$), the model simulated the natural progression of a tumor without intervention. Initially, at t_0, the tumor was localized with limited spread. Over time, the tumor expanded steadily, showing increased density and area coverage. By t_1 and t_2, the tumor had grown significantly, occupying a larger area of the simulated tissue. This scenario validated the model's ability to replicate natural tumor growth patterns in untreated conditions accurately.

The aggressive scenario ($P_{growth} = 0.075$ and $P_{metastasis} = 0.0004$) simulated conditions where factors enhance tumor growth and metastasis, with gene pathways in the Pager3 database contributing to cancer advancement activated. By t_1 and t_2, there was rapid and extensive tumor growth, with significantly increased density and area coverage compared to the control scenario. The tumor size at t_2 in the aggressive scenario was approximately 1.26 times larger than in the control scenario, demonstrating the model's ability to capture accelerated growth patterns.

The drug treatment scenario ($P_{growth} = 0.03$ and $P_{metastasis} = 0.0001$) evaluated the efficacy of therapeutic interventions in controlling tumor growth, with gene pathways in the Pager3 database contributing to cancer advancement inhibited. As time progressed to t_1 and t_2, the tumor growth was noticeably slower, and the density was reduced compared to both the Control and Aggressive scenarios. At t_2, the tumor size in the drug treatment scenario was approximately 0.78 times the size of the control scenario, illustrating the model's capacity to simulate the effects of therapeutic interventions.

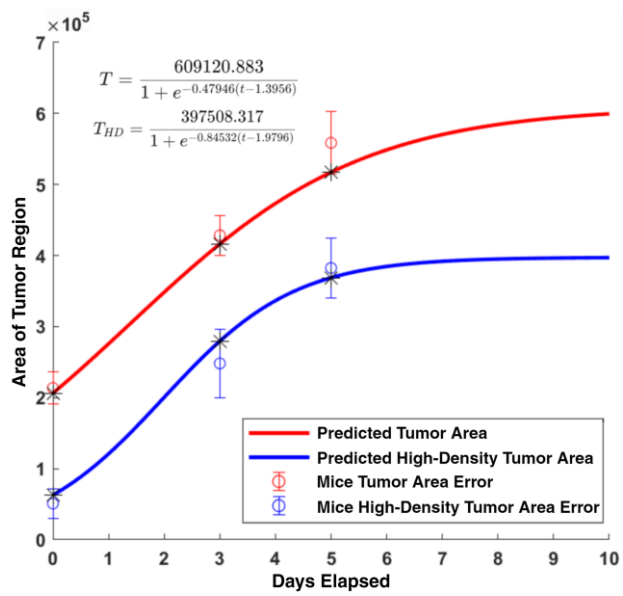

Figure 5: Sigmoid Fit of MLPA Predicted Tumor Area vs. Mice Data. This figure illustrates the comparison between the MLPA model's predicted tumor area (solid red line) and high-density tumor area (solid blue line) against the actual tumor areas observed in the mice dataset (open circles with error bars). The sigmoid functions fitted to the MLPA predictions are shown, with equations provided for both the overall tumor area $\boldsymbol{T}$ and the high-density tumor area $\boldsymbol{T_{HD}}$. The open red and blue circles represent the mean observed tumor and high-density tumor areas from the mice data at days 0, 3, and 5, respectively, with error bars indicating the range of area values.

3.2 MLPA Case Study

We wanted to test whether the MLPA model could accurately follow the cancer trajectory observed in the mice dataset, specifically on the four testing nontreated mice subjects that were not used during the training phase. MLPA was run four times, with the tumor pixel and high-density tumor pixel counts averaged. We found out that the results were able to be represented by a sigmoid function, with $N(t) = \frac{L}{1+e^{-k(t-t_0)}}$, where N(t) is the number of cancer cells at time t, L represents the carrying capacity of tumor cells, k controls the growth of the cancer, and t_0 is the time of the inflection point of the growth. The R squared value achieved was 0.93.

The sigmoid function is ideal for modeling tumor growth because it captures the initial slow growth phase, rapid exponential growth, and eventual plateau as the tumor reaches its maximum size, reflecting real-life conditions [30]. The error bars in **Fig 5** are the range of area values obtained from the testing split of the mice imaging data. Both the tumor pixel and high-density tumor predictions from MLPA fell within an acceptable range of error. MLPA predicted that the total carrying capacity would be at 1,006,630 cells, with the inflection point of growth being between 1 and 2 days. Notably, the model predicts a deceleration in cancer cell proliferation beyond Day 5, as indicated by the approaching plateau in the graph. This deceleration can be attributed to the tumor reaching spatial constraints within the host environment.

3.3 MLPA Time Point Prediction and Baseline Comparison

We aimed to test whether the MLPA model could accurately predict a time point that was not provided during the training phase, ensuring it could generate continuous, reliable data. The mice dataset was limited to only three data points, which made it insufficient for this purpose, as hiding one point would leave too few for meaningful predictions. To overcome this limitation, we used an initial testing MLPA model with the initial parameters of P_{growth} = 0.045 and $P_{metastasis}$ = 0.0003 to generate simulated imaging data for days 0, 2, 3, and 5. We then intentionally hid the day 2 cancer image and used a separate MLPA model to see if it could accurately predict the missing image without having access to the initial model's parameters, only having the other three days' imaging data to develop its own parameters (**Fig 6A**). By effectively predicting the missing data for day 2 on another model instance without knowing the initial parameters, MLPA shows its capability to generate continuous, reliable data, even when only a few time points are available (**Fig 6B**). The second model adopted the parameters to P_{growth} = 0.0456 and $P_{metastasis}$ = 0.000249 during the fitting process.

We then analyzed the tumor cell density for day 2 using the newly calibrated parameters and compared these predictions with the original day 2 observations. The 95% confidence interval for the validation MLPA model's day 2 tumor pixel count was determined to be between 239,413 and 281,988 pixels when the actual tumor pixel count was 258,758. The 95% confidence interval for the validation MLPA model's day 2 high-density tumor pixel

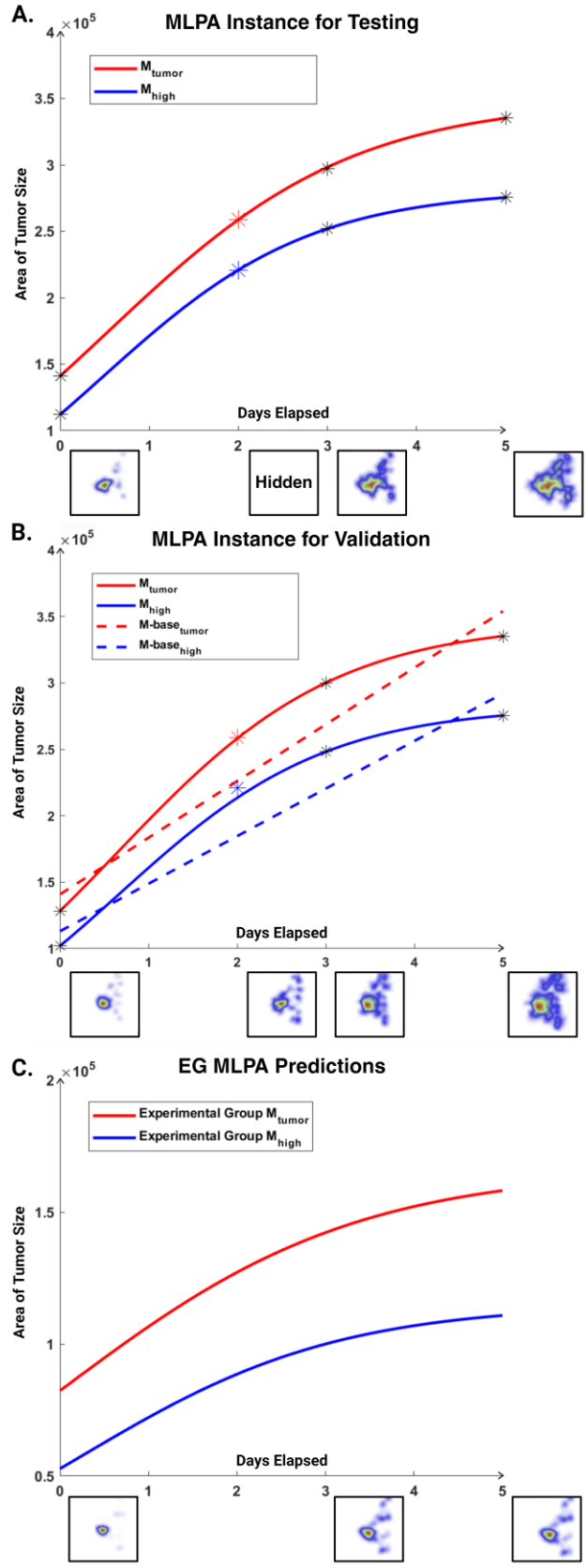

Figure 6: MLPA Model Testing, Validation, and Experimental Group Predictions. A) MLPA Instance for Testing: The tumor area (M_{tumor}) and high-density tumor area (M_{high}) for the testing MLPA model are shown over five days, with Day 2 data intentionally hidden to assess model accuracy. The asterisks represent the data points at each day, while the solid lines are the continuous MLPA model predictions. **B)** MLPA Instance for Validation: The validation MLPA model predictions for the same tumor parameters, trained on Day 0, Day 3, and Day 5 imaging data from the testing MLPA without access to Day 2 data, are shown. The dotted lines represent predictions from the baseline linear regression model ($M - base_{tumor}$ and $M - base_{high}$) for error comparison. **C)** EG MLPA Predictions: Tumor area predictions for the experimental group (4 treated mice subjects) treated with a small-molecule inhibitor of DOT1L over five days. The solid lines represent the predicted M_{tumor} and M_{high} for the experimental group. Insets below each graph show representative heatmaps from the MLPA simulations at each corresponding time point.

count was determined to be between 208,999 and 231,933 pixels when the actual high-density tumor pixel count was 220,884. In both cases, the 95% confidence interval of the validation MLPA model encapsulates the recorded testing MLPA model's pixel counts, indicating that our model's predictions for tumor cell density on day 2 are statistically consistent with the testing data.

To further validate the model's predictive capabilities, we extended our analysis by simulating tumor progression while hiding day 5 data, using only the initial parameters from days 0, 2, and 3. In this scenario, we trained a new validation MLPA model using data from days 0, 2, and 3 from the same testing MLPA model with the same model parameters as before, optimizing the parameters to best fit the observed data. The training process yielded optimized parameters of $P_{growth} = 0.0468$ and $P_{metastasis} = 0.000268$. The predicted parameters were consistent with those obtained from the validation model predicting day 2 imaging data, showing consistency when using the same data.

We then predicted the tumor cell density for day 5 using the new validation MLPA model and compared these predictions with the original day 5 observations. The 95% confidence interval for the validation MLPA model's day 5 tumor pixel count was determined to be between 324,173 and 375,250 pixels when the actual tumor pixel count was 358,011. The 95% confidence interval for the validation MLPA model's day 5 high-density tumor pixel count was determined to be between 245,823 and 287,332 pixels when the actual high-density tumor pixel count was 257,161. In both cases, the 95% confidence interval of the validation MLPA model encapsulates the recorded testing MLPA model's pixel counts, indicating that our model's predictions for tumor cell density on day 5 are statistically consistent with the original model data. This result displays the model's reliability in predicting future time points with a predetermined set of data.

We also compared our model's performance with a baseline linear regression (**Fig 6B**):

	Average Tumor Residual	Average High-Density Tumor Residual
Mtumor	666	
M-basetumor	32,710	
Mhigh		7,466
M-basehigh		36,269

These results demonstrate that MLPA's predictions for tumor and high-density cell distributions are significantly more accurate than the baseline model, effectively capturing the processes of tumor progression.

3.4 Simulating Drug Effects Through Gene Pathway Perturbation

Four mice with MLL-rearranged leukemia in our study were treated with a small-molecule inhibitor of DOT1L. The efficacy of this treatment was assessed by its ability to revert gene perturbations caused by the disease. Instead of changing the default parameters of the MLPA model, we selected specific gene pathways known to be inhibited by DOT1L, which were then used to perturb the default model parameters of $P_{growth} = 0.055$ and $P_{metastasis} = 0.0002$ (**Fig 6C**). We selected the following pathways from our gene pathway file which would be inhibited during treatment [29, 30]:

1. Nonsense Mediated Decay
2. Chaperon Mediated Autophagy
3. Regulation of TP53 Expression and Degradation
4. TP53 Activity through Methylation
5. Aggrephagy

These pathways were chosen for their roles in various cellular processes that can counteract cancer progression, such as degrading potentially oncogenic proteins, maintaining protein homeostasis, enhancing tumor suppressor functions, and reducing proteotoxic stress.

We created a 95% confidence interval for the day 5 experimental group predicted tumor pixel counts and compared it to the day 5 pixel counts of the drug treated mice in our dataset. The 95% confidence interval for tumor pixel count was determined to be between 153,336 and 170,231 pixels when the actual tumor pixel count was 158,125. The 95% confidence interval for high-density tumor pixel count was determined to be between 97,215 and 120,958 pixels when the actual tumor pixel count was 110,807.

To further validate these results, we performed equivalence tests using the Two One-Sided Test (TOST) procedure. For tumor pixels, the equivalence margin was set at ±3,953 pixels (2.5% of the observed value), with a 95% simulated predictive interval of [143,124, 167,784]. The difference between MLPA's and the imaging dataset means was 2,670 pixels. For high-density tumor pixels, the equivalence margin was ±2,770 pixels (2.5% of the observed value), with a 95% simulated predictive interval of [101,545, 118,347]. The difference between MLPA's and the imaging dataset means was 860 pixels.

In both cases, MLPA's predicted pixel counts fell within the intervals from the actual imaging dataset and with differences well

within the defined equivalence margins. These findings indicate that the drug treatment significantly reduced tumor size and density, aligning closely with our model's predictions. The strong agreement within the confidence intervals underscores the model's ability to accurately capture the effects of activated gene pathways, thereby validating our approach to simulating drug effects through gene pathway perturbation. This alignment not only reinforces the model's predictive accuracy but also supports its potential utility in personalized medicine applications where understanding the impact of gene-targeted therapies is crucial.

4 Conclusion

Developing the MLPA framework marks a significant step forward in simulating cancer growth, metastasis, and treatment responses by integrating macroscopic and microscopic biological data. Our results demonstrate MLPA's capability to accurately model tumor progression under various conditions and predict the effects of targeted therapies.

The MLPA model showed high fidelity in simulating different scenarios of tumor growth, as evidenced by the control, aggressive, and drug treatment simulations (Section 3.1). The model's ability to capture the natural progression of untreated tumors, the accelerated growth in aggressive conditions, and the reduced growth under drug treatment demonstrates its versatility and potential utility in clinical settings.

Using bioluminescent imaging data from mice, our validation process showed that MLPA can accurately fit and predict tumor growth patterns. The sigmoid function fitting (Section 3.2) captured tumors' characteristic S-shaped growth curve, including the initial slow growth, rapid proliferation, and eventual plateau phases. This aligns well with established tumor growth models and biological observations.

The predictive accuracy of MLPA, as demonstrated in Section 3.3, is particularly noteworthy. The model's ability to predict tumor cell density on day 2, using only data from days 0, 3, and 5, with high statistical consistency, speaks to its robustness. The significant improvement in prediction accuracy compared to baseline linear regression (error reduction by a factor of 49 for tumor pixels and 4.9 for high-density tumor pixels) underscores the power of MLPA's multi-scale, stochastic approach.

Furthermore, MLPA's capacity to simulate drug effects through gene pathway perturbation (Section 3.4) represents a significant strength of the framework. The close alignment between predicted and observed tumor sizes in treated mice, validated through equivalence testing, suggests that MLPA can accurately model the complex interactions between drug treatments and tumor biology at a molecular level.

Compared to existing cancer modeling approaches, MLPA offers several advantages. The first is multi-scale integration. MLPA uniquely combines macroscopic (EHR) and microscopic (genomic) data, providing a more comprehensive representation of tumor dynamics than models focusing on a single scale. The second is flexibility. The stochastic cellular automata approach allows MLPA to simulate a wide range of scenarios, from natural tumor progression to the effects of various treatments. The third is predictive accuracy. Our results demonstrate that MLPA shows high predictive accuracy, outperforming baseline regression models and accurately capturing the effects of simulated drug treatments. Fourth and last is biological relevance. By incorporating gene pathway information, MLPA provides insights into the molecular mechanisms underlying tumor growth and treatment response, a feature often lacking in purely statistical or phenomenological models.

However, it is essential to acknowledge the limitations of our current approach. The model's validation was performed using mouse data, and while this provides a strong proof of concept, further validation with human cancer data is necessary to confirm its clinical applicability. Additionally, the current implementation is limited to 2D simulations, which, while computationally efficient, may not capture all aspects of 3D tumor growth in vivo.

5 Discussion

MLPA represents a promising advance in personalized cancer modeling. It offers a unique combination of multi-scale data integration, predictive accuracy, and biological relevance. The framework shows great potential as a tool for optimizing cancer treatment strategies and advancing our understanding of tumor biology.

Future developments are essential to further enhance the model's accuracy and applicability. One area for improvement involves expanding the capabilities of our Monte Carlo simulations. Currently, as we are visually graphing these simulations, the model can manage up to three variables in 3D space, limiting their complexity. Future research should explore advanced computational techniques, such as parallel computing or high-performance computing resources, to increase the efficiency and scalability of these simulations.

One of the key limitations of our current approach is the lack of paired genomic and imaging data in the mice dataset, which restricts our ability to fully validate the extent to which specific gene pathways perturb the model's default parameters. While our framework can integrate genomic data to refine model parameters, the precise values used in this study were set based on assumptions and may not be generalizable across different datasets. Future work will focus on obtaining datasets that include both imaging and genomic information, allowing for a more accurate and data-driven customization of the model. This will enable more precise predictions of tumor behavior and treatment responses, enhancing the clinical relevance of the MLPA framework. In addition, while the MLPA framework is designed to incorporate EHR data, such as patient weight, age, gender, blood pressure, and height, to personalize the model parameters further, this capability was not fully evaluated in the current study. The use of mice imaging data, which did not include EHR data, and the incompatibility between animal EHR data and human EHR data, limited our ability to validate the impact of these factors. Future research should focus on integrating and testing human EHR data to explore how these

additional physiological parameters influence the model's predictions and overall effectiveness in a clinical setting.

While further validation and development are needed, particularly with human cancer data, MLPA shows great promise in advancing personalized medicine approaches for cancer treatment. Future work will focus on addressing the current limitations and expanding the model's capabilities, aiming to improve patient outcomes through more precise and personalized cancer therapies. The digital twin model will continue to be a valuable tool in simulating a broader range of biological processes and therapeutic scenarios, offering tailored and effective treatments that significantly improve patient outcomes.

ACKNOWLEDGMENTS

Both authors acknowledge the generous support of the AlphaMind Club, UAB research startup fund, and NIH grant award 1OT2:OD032742-01 to JYC.

REFERENCES

[1] Vogelstein, B., & Kinzler, K. W. (2004). Cancer genes and the pathways they control. Nature Medicine, 10(8), 789-799.
[2] Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. Nature, 458(7239), 719-724.
[3] Caldas, C., & Brenton, J. D. (2005). Sizing up miRNAs as cancer genes. Nature Medicine, 11(7), 712-714.
[4] I. Dagogo-Jack, & A. T. Shaw, 2018. Tumour heterogeneity and resistance to cancer therapies. Nature Reviews Clinical Oncology, 15(2), 81-94.
[5] D. Hanahan, & R. A. Weinberg, 2011. Hallmarks of cancer: the next generation. Cell, 144(5), 646-674.
[6] A. Conde-Rodríguez, C. Sarabia-Cobo, & A. González-Vidal, 2022. Digital twins in healthcare: A systematic literature review. Journal of Medical Systems, 46(8), 1-14.
[7] Altrock, P. M., Liu, L. L., & Michor, F. (2015). The mathematics of cancer: integrating quantitative models. Nature Reviews Cancer, 15(12), 730-745.
[8] C. A. Valentim, J. A. Rabi, S. A. David, 2023. Cellular-automaton model for tumor growth dynamics: Virtualization of different scenarios. Computers in Biology and Medicine, 153
[9] P. Mascheroni, S. Savvopoulos, J. C. L. Alfonso, M. Meyer-Hermann, & H. Hatzikirou (2021). Improving personalized tumor growth predictions using a Bayesian combination of mechanistic modeling and machine learning. Communications Meidcine, 19
[10] S. Turajlic, A. Sottoriva, T. Graham, & C. Swanton, 2019. Resolving genetic heterogeneity in cancer. Nature Reviews Genetics, 20(7), 404-416.
[11] G. P. Gupta, & J. Massagué, 2006. Cancer metastasis: building a framework. Cell, 127(4), 679-695.
[12] R. Laubenbacher, J. P. Sluka, & J. A. Glazier, 2022. Using digital twins in viral infection. Science, 371(6534), 1105-1106.
[13] Swedish Digital Twin Consortium. (n.d.). Retrieved from https://www.sdtc.se/
[14] Gu, A. (2018). CMM Perturbation Theory and AI Engineering Design Optimization. ResearchGate, 1-12.
[15] J. Metzcar, Y. Wang, R. Heiland, & P. Macklin, 2019. A review of cell-based computational modeling in cancer biology. JCO Clinical Cancer Informatics, 3, 1-13.
[16] A. Karolak, D. A. Markov, L. J. McCawley, K. A. Rejniak, 2018. Towards personalized computational oncology: from spatial models of tumour spheroids to organoids to tissues. Journal of the Royal Society Interface, 15(138), 20170703.
[17] Sun, H., Pisle, S., Gardner, E. R., & Figg, W. D. (2010). Bioluminescent imaging study: FAK inhibitor PF-562271 preclinical study in PC3M-luc-C6 local implant and metastasis xenograft models. Cancer Biology & Therapy, 10(3), 235-243.
[18] K. Tiwari et al (2023). ChatGPT usage in the Reactome curation process.
[19] Björnsson, B., Borrebaeck, C., Elander, N., Gasslander, T., Gawel, D. R., Gustafsson, M., Jörnsten, R., Lee, E. J., Li, X., Lilja, S., Martínez-Enguita, D., Matejka, A., Sandberg, O., Fryknas, M., Jirstrand, M., Benson, M. (2020). Digital twins to personalize medicine. Genome Medicine, 12(1), 4.
[20] Lim, S. L., Tan, C. S., Chua, J. W., & Teo, C. (2015). Integrating physiological data into personalized cancer management: Opportunities and challenges. Annals of the Academy of Medicine Singapore, 44(10), 361-368.
[21] Armitage, P., & Doll, R. (2004). The age distribution of cancer and a multi-stage theory of carcinogenesis. British Journal of Cancer, 91(12), 1983-1989.
[22] Li, T., Fan, J., & Wang, B. (2016). TIMER: a web server for comprehensive analysis of tumor-infiltrating immune cells. Cancer Research, 77(21), e108-e110.
[23] Kfoury, M., Moubarak, M., & El-Khoury, C. (2021). The Influence of Age and Gender on Cancer Progression: State-of-the-Art and Implications for Personalized Treatment. Cancers, 13(15), 3918.
[24] Poleszczuk, J., & Enderling, H. (2016). A High-Performance Cellular Automaton Model of Tumor Growth with Dynamically Growing Domains. Applied Mathematics, 7(14), 1665-1682.
[25] S. Wolfram, 2024. Why does biological evolution work? A minimal model for biological evolution and other adaptive processes. Retrieved from https://writings.stephenwolfram.com/2024/05/why-does-biological-evolution-work-a-minimal-model-for-biological-evolution-and-other-
[26] Hu, J., Gao, Y., Li, J., Shang, X., & Wang, J. (2021). PAGER 3.0: an update of the pathway annotated-list and gene-signature electronic repository for Human Network Biology. Nucleic Acids Research, 49(D1), D1347-D1354.
[27] Wang, E., Kawaoka, S., Yu, G., Xu, Y., Kurosu, T., & Tanaka, Y. (2013). Histone H2B ubiquitin ligase RNF20 is required for MLL-rearranged leukemia. Proceedings of the National Academy of Sciences, 110(11), 4758-4763. DOI: https://doi.org/10.1073/pnas.1218838110.
[28] I. Harshe, H. Enderling, R. Brady-Nicholls, (2023). Predicting Patient-Specific Tumor Dynamics: How Many Measurements Are Necessary? Cancers (Basel) 15(5):1368.
[29] Harrison, R. L. (2010). Introduction To Monte Carlo Simulation. AIP Conference Proceedings, 1204(1), 17-21.
[30] Kozusko, F., Bourdeau, M. A unified model of sigmoid tumour growth based on cell proliferation and quiescence. Cell Prolif. 2007 Dec;40(6):824-34.