“Modelling of Disease Progression in Myeloproliferative Neoplasms”

„Modellierung der Progredienz von Myeloproliferativen Neoplasien“

Von der Fakultät für Maschinenwesen der Rheinisch-Westfälischen Technischen Hochschule Aachen zur Erlangung des akademischen Grades einer Doktorin der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Maryam Montazeri Ghahjavarestani

Berichter:	Universitätsprofessor Dr. rer. nat. Andreas Schuppert
	Universitätsprofessor Dr. med. Steffen Koschmieder
Tag der mündlichen Prüfung:	05.11.2019

Abstract

This study focuses on the modeling of the disease progression in Myeloproliferative Neoplasms (MPNs), which are a group of chronic blood cancers with the possibility of leukemic transformation. For modeling disease progression in MPNs and by considering cancer progression as a multistep process which engages a hierarchy of different levels of functionalities, we developed hybrid models using datasets from heterogeneous sources such as gene expression, simulated data from population dynamics models and clinical laboratory results.

Based on the different genetic origin of different MPNs, we divided our research into two main parts: Philadelphia positive MPNs and Philadelphia negative MPNs.

In the first part and for the Philadelphia positive MPN, with the help of systems theory concepts, we developed a hybrid model by integrating data-driven models and a population-based-mechanistic model.

In the second part and for the Philadelphia negative MPNs, we developed a hybrid classification model by the integration of smaller sub-models on separate subspaces of the available feature space.

We showed that the use of the concept of hybrid modeling for tracking disease progression in the Philadelphia positive MPN enables patient-specific risk assessment to avoid leukemic transformation.

Based on our results in the Philadelphia negative MPN part, we showed that hybrid models not only result in more accurate classification and diagnosis of MPNs but also help reduction of overfitting problem of pure data-driven models. Our developed models for MPNs classification can be improved by the integration of more data types such as omics data, which is at the current time unfortunately not available. More data on different levels can add more information about the involved mechanisms at different levels of cancer progression hierarchy.

Zusammenfassung

Myeloproliferative Neoplasien (MPN) sind eine Gruppe von malignen chronischen Blut-Erkrankungen, in deren Verlauf eine Leukämie entwickelt werden kann. Ziel dieser Dissertation war es, das Fortschreiten dieser Blut-Krankheiten als mehrstufigen Prozess zu modellieren, der das komplexe Zusammenspiel der verschiedenen beteiligten biologischen Instanzen und deren ihrer Hierarchie entsprechenden spezielle Funktionalitäten berücksichtigt. Mithilfe der Verknüpfung von heterogenen Datensätzen, wie zum Beispiel gemessenen Genexpressionsdaten, Populationsdynamik Simulationen und klinischen Laborerkenntnissen, wurden dazu im Zuge dieser Arbeit hybride Modelle entwickelt.

Basierend auf der genetischen Zusammensetzung der verschiedenen MPNs, ist diese Arbeit in zwei Teile gegliedert: Philadelphia-positive MPN und Philadelphia-negative MPN. Der erste Teil umfasst die Entwicklung eines hybriden Modells für Philadelphia-positive MPN, das mithilfe von Konzepten der Systemtheorie einen datengetriebenen Modellierungsansatz mit einem populationsbasierten mechanistischen Modell kombiniert. Für die Philadelphia-negative MPN im zweiten Teil der Arbeit wurde hingegen ein hybrides Klassifikations-Modell entwickelt, das mehrere kleine Submodelle aus separaten Modell-Unterräumen des verfügbaren Merkmalraumes integriert.

Mit den im ersten Teil dieser Arbeit vorgestellten Untersuchungen wird gezeigt, dass das Konzept der hybriden Modellierung zur Erfassung der Progredienz der Philadelphia-positive MPN eine patienten-spezifische Risikobewertung ermöglicht, mithilfe derer eine leukämische Transformation der Krankheit vermieden werden kann. Die Ergebnisse aus dem zweiten Teil dieser Arbeit im Zusammenhang mit Philadelphia-negativen MPN demonstrieren, dass hybride Modelle im Gegensatz zu ausschließlich daten-getriebenen Modellen helfen Probleme wie Overfitting zu reduzieren und die Klassifizierung und Diagnose von MPNs verbessern können.

Die Resultate dieser Arbeit deuten an, dass eine Eingliederung quantitativer Daten von verschiedenen funktionellen Ebenen die Prognose-Modelle verbessern und enorm zum weiteren Verständnis der verschiedenen Mechanismen in der Krebsprogression beitragen könnten. Aufgrund der Limitierung durch die zur Verfügung stehenden Datentypen beschränken sich die Analysen dieser Arbeit jedoch auf Genexpressionsmodellierung und verzichten auf die Integration anderer Omics-Datentypen.

Acknowledgments

It is a great chance here to thank, first of all, my supervisor, Prof. Andreas Schuppert, for supporting me during my doctoral studies with his great knowledge, valuable advice, and of course, his patience. I learned a lot from him not only from the scientific point of view but also from his personality and manner. I learned from him to take the chance of learning from everything, to be humble, and to widen my vision.

I also want to thank Prof. Steffen Koschmieder for supporting me with his great knowledge in the field of hematology, helping me to have a better understanding of myeloproliferative neoplasms and to come up with new ideas.

I am thankful to my colleagues Ali Hadizade and Pejman Farhadi for the fruitful discussion I had with them and their creative ideas. I also want to thank Lisa Schätzle for helping me in translating the abstract of this document in German, and being a supportive friend.

I thank all my colleagues in JRC-Combine for all the scientific discussions and fun times I had with them in several retreats. I will miss JRC, and the time I spent with all of you

I am also very grateful to Hülya Ulu-Esser for her administrative support.

A special big thank goes to my family in Iran for their unconditional love and support through all the difficult times I had during my doctoral studies. I am also grateful to my brother, Alireza, for proofreading of this document.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1-1: MOTIVATION AND GOAL OF THE CURRENT STUDY

Patient-specific disease progression models have recently gained a lot of interest in the research communities in order to do clinical decision makings based on the individual characteristic profile of each patient rather than on the average patient model. There are so many benefits of personalized medicine such as planning more effective medication strategies, lower adverse event risk incidence rate, and lower healthcare costs to name a few[1]. In our study, we try to apply personalized medicine strategy in the field of modeling phase transition in chronic diseases, especially in cancer. The high complexity of cancers and interference of different factors in the evolution of cancer makes it extremely difficult for the exact patient-specific prognosis based on all possible covariates. Hanahan and Weinberg[2] proposed common hallmarks across all cancers, which provide a logical framework for understanding how healthy cells evolve progressively to a neoplastic state by passing through a multistep process. So, instead of fully understanding the complexity of cancer we can focus on detecting the phase transition points for more effective clinical management. Therefore, by considering cancer as a complex process across multiple scales and levels of functionalities and by looking at the formation of malignancy of cancer from a benign onset through *transition points*, we try to introduce and apply techniques and methods, which can model these phase transitions in human cancer in a multi-level approach. To achieve the primary purpose of our study, we focus on Myeloproliferative Neoplasm (MPNs). MPNs are blood disorders with some specific features, which help in modeling the phase transition in them. First of all, since they follow multistep carcinogenesis[3] as a developing process, which eventually leads to the formation of heterogenic clones, they can be served as a model for investigating the transition patterns. What makes them ideal to start the modeling process with, is their less complexity comparing to solid tumors such as lung or liver cancers. Additionally, MPNs are blood cancers, which have been the subject of many research and studies focusing on the understanding of underlying molecular mechanisms triggering such cancers. As studies suggest, the onset of MPNs is usually rooted in a single mutation in stem cells. However, this benign onset of MPN subtypes is eventually followed by leukemic transformations or myelofibrosis. This simplicity at the onset stage of the disease is another feature of MPNs, which makes them unique to model the disease progression. Also, thanks to clinical monitoring techniques, unlike some other cancers such as the liver and pancreatic cancers, which are identifiable only in later stages of the disease, MPNs can be robustly identified in early clinical stages. However, modeling disease progression in MPNs ultimately suffers from the general problems that challenge modeling most of the other human diseases. We tried in this study to identify patterns for tracking the MPN progression and for the detection of possible phase transitions. Modeling cancer progression in the human body needs overcoming specific challenges that are not present when dealing with other systems such as fluid mechanics, physics, chemical engineering, etc. Heterogeneity of cells, the complexity of mutual interactions, presence of different enzymes and hormones, different mesh conditions between *in-vitro* and *in-vivo* environment are to name a few of these challenges.[4] With all that being said, the main objective of this research is to track the disease progression in MPNs by identifying biomarkers that can capture the onset of the phase transition from a benign course of the disease to a malignant one in each individual patient. The prediction of the onset of phase transition has high clinical

relevance and is helpful for better clinical decision makings, such as better diagnostic and therapeutic management.

1-2: STATE OF THE ART AND THE NEED FOR FURTHER STUDIES

Most of the studies aiming for modeling MPNs focus either on simulation of cell population dynamics purely through low complex mechanistic models or on developing machine learning models on images of chronic myeloid leukemia (CML) bone marrow for better classification of different CML stages. For the review of available mechanistic and data-driven models in MPNs please refer to chapter two. Available mechanistic models focus on the dynamics of blood cell populations as a result of the occurrence of mutations in hematopoietic stem cells. Here it is also worth mentioning that mechanistic models in MPN research are limited on CML because of the simpler driving mechanism of CML progression due to a specific mutation, whereas for other MPNs, up to date, such mechanistic models do not exist. These mechanistic models, however, mostly represent a core disease model based on key drivers of the disease in an average patient and do not consider the individual characteristic profile of patients[1]. Therefore, they cannot provide any patient-specific prediction of phase transition. Disease progression in chronic diseases is a result of a hierarchy of different functionalities in different scales[5], and therefore, a hierarchy of models in different functionality levels is needed[6]. To date, there are no models that consider all levels and scales, such as changes in genetic makeup, dynamics of cell populations, and patient-specific confounders such as co-morbidities, patient history, and physiological conditions.

1-3: PRESENTED SOLUTIONS IN THE CURRENT STUDY

To overcome the mentioned challenges, we have designed our study to benefit from Integration of the heterogeneous data sources such as gene expression datasets, population dynamics models, and clinical laboratory test results in order to consider possibly most of the different levels of the disease progression hierarchy.

Gene expression data sources have been used in so many studies so far and are useful in capturing genetic variations between healthy and control patients. We assumed that these variations in gene expression profiles are somehow the result of changes in cell types populations over the time course of disease progression. Therefore, dynamic cell population models or cell counts data in laboratory test results as an alternative to population dynamics models, can benefit us in confronting the challenges arising because of heterogeneity of cells. Coupling gene expression data and population dynamics models help us to captivate the genotypic and phenotypic changes, which are mirrored only in one data type and are absent in the other one. It also ultimately leads to develop a robust integrated biomarker to track the disease progression in each individual patient. Due to the different genetical raise of different MPNs, we divided our study into two main panels: the first panel focuses on tracking disease transition in CML by the integration of existing mechanistic models and patients' gene expression data. In another panel, we mainly focus on the classification of Philadelphia negative MPNs based on gene expression data in one section and by developing hybrid models on clinical data in another section.

In the CML panel, we present a novel method for tracking disease progression and prediction of the critical phase transition point by integrating two different levels of data. Firstly, cell population dynamics data from existing mechanistic models are used for the core disease model in an average patient, and secondly, gene expression data of CML patients are used to

complete our patient-specific modeling approach. To be able to integrate these different levels of data, we applied well-established methods of identifying critical states in general complex systems such as increased entropy. The workflow of this model, therefore, consists of the following steps (figure 17 in section one of chapter 4):

- Introducing a biomarker which is accessible on both sides of the integration model, i.e. simulated cell-population data and patients' gene expression
- Assessment/calculation of patient-derived biomarker for chronic CML progression from gene expression data
- Assessment/calculation of corresponding model-derived markers from cell population data
- Introducing an alignment marker which is assessable in both sides of our integration model.
- Mapping of the patients with the help of entropy markers, to the time axis of the mechanistic model

With the help of this model, we are able to locate each individual patient on the time spectrum of disease progression in CML chronic phase. The rationale and workflow of the model are fully described in chapter 4.

In the MPN panel, we don't have access to any mechanistic models simulating the dynamics of cell populations, therefore, as an alternative, we used cell count data of MPN patients from laboratory results in one level of our hierarchical modeling strategy. In another level, we used other data types such as patient history, co-morbidities data, therapy information, and physiological status. The integration of these two-level data helps to develop a hybrid model, which enables us not only to overcome the curse of dimensionalities which often emerge as a problem of developing data-driven models with large input space but also to make better diagnostic decisions in clinics (figure 42 in section 2 of chapter 5). The workflow of the MPN classification model is fully described in chapter 5, however here we shortly introduce major steps:

- Division of the data space into two parts: cell-count data and non-cell count data
- Setting up a classifier based on the cell-count data
- Setting up a regression model for predicting the classification error of the cell-count based model
- Integration of the classifier and the regression model to justify the classification accuracy

To develop this hybrid model, we have tried different available computational models to select the best one with the highest diagnostic accuracy. We Also discovered genetic differences between different stages of classical MPNs to extract a disease progression biomarker which locates different MPN patients in a spectrum of disease progression from early stage to more malignant stages. With the help of this biomarker, we then further analyzed the progression of MPNs across different blood cell types or blood-related tissues utilizing the previously established models which extract physiological relevance out of MPN patients' gene expression profile.

1-4 OUTLINE OF THE CURRENT STUDY

The outline of this work is described as follows: In the second chapter, we tried to give a background on MPNs in two main divisions: CML and Philadelphia-negative MPNs. We provide a summary of the genetical, phenotypical, and morphological features of MPNs and the developed risk scoring systems. Afterward, we review shortly state of the art in mathematical modeling concepts of disease progression in CML and MPNs separately. In chapter three and the method section, we provide conceptual basics of the methods we used in the study, such as hybrid modeling, systems theory, and machine learning algorithms. In the materials section, we prepared a list of data used in this study. In chapter four, we introduce a novel approach for modeling disease progression in chronic CML through the integration of data-driven and mechanistic markers. We show that our approach can shed light on detecting phase transition in chronic CML. In Chapter five, we investigate Philadelphia negative MPNs in sperate levels of data types i.e., gene expression and clinical data. In the gene expression analysis, in section 5-2, we introduce a biomarker capable of tracking disease progression and classifying different stages of MPN. In the clinical data level, in section 5-3, we developed a hybrid model for a better classification of MPN stages. In chapter 6, with the help of the thermodynamic concept of entropy, we show how the heterogeneity of different levels of cell population and gene expression are related together. Chapter seven presents a summary of our study. We also provide an appendix where all the details of methods and materials are described for reproducing the results.

CHAPTER 2: INTRODUCTION ON MPNS

This introduction is a literature review on MPN research and the challenges of studying these kinds of blood disorders, as well as state of the art in modeling critical states in these types of blood cancers. MPNs, in a broad view and based on the genetic basis, are divided into two major categories: CML or Philadelphia positive MPN and Philadelphia negative MPNs. This introduction consists of two major parts: the background of CML research and the background of MPN research.

2-1: PHILADELPHIA POSITIVE MPN OR CHRONIC MYELOID LEUKEMIA

In this section, it has been tried to give an overview of the CML genomic origin, phenotypes, symptoms, models for risk stratification in CML, existing therapy options for CML, and the related works on modeling disease progression in CML.

2-1-1: GENETIC BASIS:

Chronic myeloid leukemia (CML) is a hematopoietic stem cell (HSC) disorder caused by the formation of Philadelphia chromosome or BCR-ABL chromosome which is the result of a translocation of chromosomes 9 and 22.[7,8]

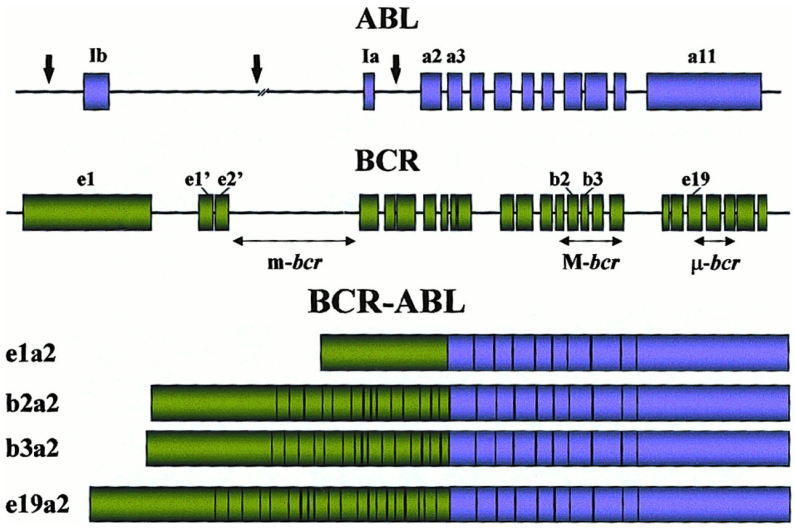


FIGURE 1:LOCATIONS OF THE BREAKPOINTS IN THE ABL AND BCR GENES AND STRUCTURE OF THE CHIMERIC MRNA OF DERIVED FROM VARIOUS BREAKS. THE FIGURE IS TAKEN FROM[9].

The occurrence of BCR-AbL1 oncogene in a single HSC affects its apoptotic potential, cell division rate, and differentiation capacity when compared to normal HSCs[10]. BCR-ABL1 interrupts several cell functionalities and signaling pathways, such as deregulation of the Abl tyrosine kinase, altering cell adhesion properties, activation of mitogenic signalings such as Ras, and MAP kinase pathways, JAK-STAT pathway, PI3 kinase pathway, MYc pathway inhibition of apoptosis, and degradation of inhibitory proteins[9]. Philadelphia chromosome is present in all CML patients and therefore, it serves as a biomarker for diagnosis, and a susceptible target for therapy[11].

2-1-2: CML EVOLUTION AND DIAGNOSIS

CML is classically staged into three phases, chronic phase (CP), accelerated phase (AP), and blast phase (BP).[12] In another word, CML evolves from CP to AP to BP if the patient receives no effective therapy in the early stages. Chronic CML patients are mostly diagnosed based on abnormal blood counts[13] and showing symptoms such as anemia, splenomegaly, fatigue, weight loss, anorexia, thrombosis, and bleeding from thrombocytopenia or platelet dysfunction[14]. CML staging mostly is based on the number of blasts in peripheral blood and bone marrow and abnormalities in the number of basophil and platelets and spleen size. The World Health Organization (WHO)[12] and European leukemia Net (ELN) [15] criteria for staging chronic myeloid leukemias brought in table 1.

TABLE 1: WHO AND ELN CRITERIA FOR CML STAGING

	WHO criteria	European Leukemia Net criteria
Chronic phase	**None of the criteria for AP and BC**	
Accelerated Phase:		
Blasts in peripheral blood or bone marrow	$10-19\%$	$15-29\%$ *or blasts plus promyelocytes in peripheral blood or bone marrow* $>30\%$ *with blast* $<30\%$
Basophils in peripheral blood	$\geq 20\%$	$\geq 20\%$
Platelets	$\leq 100 \times \frac{10^9}{L}$ *not attributable to treatment, or* $plateltes > 1000 \times 10^9$ *uncontrolled on treatment*	$\leq 100 \times \frac{10^9}{L}$ *not attributable to treatment*
Additional Chromosomal abnormalities	*Occurring on treatment*	*Occurring on treatment*
White cell count and spleen size	*Increasing and uncontrolled on treatment*	…
Blast Crisis		
Blast in peripheral blood or bone marrow	$\geq 20\%$	$\geq 30\%$
Blast proliferation	*Extramedullary, except the spleen*	
Large foci of blasts	*Bone marrow or spleen*	*Extramedullary, except the spleen* …

2-1-2A: HISTOLOGY OF CML

Bone marrow and peripheral blood histology can reveal different stages of CML and can be assessed for a better diagnosis. Leukocytosis, together with circulating immature myeloid cells, is manifested in pictures of peripheral blood. In pictures of bone marrow, the presence of blast and promyelocytes are considerably much more notable (figure 2).

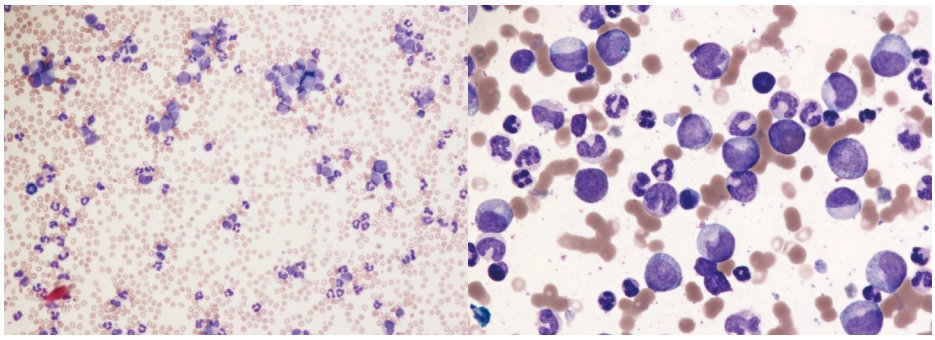


FIGURE 2: PICTURES OF PERIPHERAL (LEFT) AND BONE MARROW (RIGHT) FROM CP-CML). THE FIGURE IS TAKEN FROM [14]

Shararma and colleagues [16] investigated the histological changes of bone marrow in CP-CML and BC-CML. Figure 3 shows the results of their study. Panmyelosis and absence of adipocytes are observable in the hypercellular marrow of the chronic phase (figure 3a). In the blast phase, immature cells form clusters that fill the entire in-trabecular space (figure 3b).

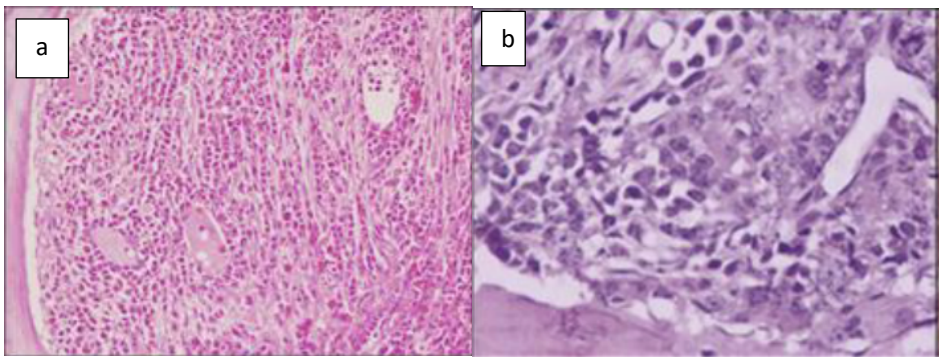


FIGURE 3:PICTURES OF BONE MARROW IN CML, IN CHRONIC PHASE(A) AND BLAST CRISIS (B). THE FIGURE IS TAKEN FROM[16].

2-1-3: RISK STRATIFICATION FOR PATIENT SURVIVAL-STATE OF THE ART

Meanwhile, there are some scoring systems such as Sokal, Hasford, and EUTOS developed for the prediction of survival in CML patients. Sokal score [17]is developed for patients treated with busulfan and is calculated for patients aged 5-48 years by the following equation:

$$riskscore = exp\ 0.0116 \times (age - 43) + 0.345 \times (spleen\ cm\ below\ costal\ margin - 7.5cm) + 0.188 \times ((\frac{plateletcount}{700})^2 - 0.563) + 0.0887 \times (blast\ percentage - 2.1) \qquad \textbf{Eq. (1)}$$

Depending on the calculated value, CML patients can be placed in three risk groups:

- low risk for scores < 0.8
- intermediate risk for 0.8<scores<1.2
- high risk for scores>1.2

Hasford score[18] is another risk stratification scoring system, which was developed for the patients treated with interferon α and is calculated based on the following equation:

$$risk\ score = (0.666 \times age(0\ when < 50\ old\ or\ 1\ otherwise) + 0.042 \times spleen\ size\ in\ cm\ below\ costal\ margine + 0.0586 \times\ blast\ percentage + 0.2039 \times basophile(0\ when < 3\%\ or\ 1\ otherwise\ ) + 1.0956 \times platelets\ count\ (0\ when < 1500\ , 1\ otherwise)) \times 100 \quad \textbf{Eq. (2)}$$

According to the calculated score, patients are placed in three risk groups:

- Low risk ($scores \leq 780$)
- intermediate risk($780 < score \leq 1480$)
- high risk for ($scores > 1480$)

ETOUS score [19] is the most recently developed score for risk assessment of CML patients and is derived based on a study on patients who are treated with tyrosine kinase inhibitor (TKI). ETOUS score is calculated according to the following equation:

$$risk\ score = basiphiles\ in\ percentage\ \times 7 + spleen\ size\ in\ cm\ below\ costal\ margin\ \times 4 \quad \textbf{Eq. (3)}$$

And patients are placed either in a low-risk category with scores ≤ 87 or in a high-risk category with scores > 87.

2-1-4: TREATMENTS

Before the discovery of the Philadelphia chromosome as the primary driver of CML and the beginning of the research for developing target therapies, there were different ways of treating CML patients like arsenic trioxide for induction of apoptosis of cancer cells[22], and splenic radiation to reduce the degree of splenomegaly[23]. Laters, Busulfan showed to be correlated with better survival of CML patients[23], and some studies indicated hydroxycarbamide, interferon$-\alpha$ [24,25,26,27, 28] as an effective therapy option. Allogenic stem cell transplantation[29] was also another option for CML patients in the chronic and accelerated phases before the emergence of tyrosine kinase inhibitors (TKIs).

Since1996 that Druker and colleagues [30] reported success in both hematological and cytogenic remission in response to the use of imatinib in-vitro, imatinib had revolutionized the management of CML [31]. There are also second-generation TKIs such as dasatinib, nilotinib, and bosutinib and recently third-generation ponatinib. However, imatinib has remained as the most popular TKI I in treating CML patients[11]. However, like other drugs, TKI also may cause some adverse side effects. A summary of available TKIs for the treatment of CML patients and possible side effects are brought in table 2 from [14].

TABLE 2: TKI AVAILABLE FOR THE TREATMENT OF CP-CML AND THEIR ADVERSE EFFECTS

Tyrosine kinase inhibitor	Common adverse effects (> 10% patients)	Notable adverse events
Imatinib	Myelosuppression, fatigue, insomnia, depression, dizziness, fluid retention, weight gain, upper respiratory tract infection, influenza, pyrexia, cough, abdominal, pain, nausea, vomiting, diarrhea, myalgia, arthralgia, skin rash, hemorrhage	Rarely associated with appendiceal carcinoma [20]
Dasatinib	Myelosuppression, fatigue, fluid retention, headache, dyspnea, infection, abdominal pain, nausea, diarrhea, myalgia, arthralgia, skin rash, hemorrhage	Pleural effusion
Nilotinib	Myelosuppression, fatigue, nausea, vomiting, diarrhea, constipation skin rash pruritis	QT interval prolongation
Bosutinib	Myelosuppression, fatigue, dyspnea, cough, pyrexia, abdominal pain, nausea, vomiting, diarrhea, constipation, increased aspartate and alanine aminotransferase, arthralgia, skin rash	
Ponatinib	Myelosuppression, fatigue, headache, hypertension, dyspnea, pyrexia, abdominal pain, pancreatitis, nausea, myalgia, arthralgia, skin rash, dry skin, increased alanine aminotransferase	Arterial and venous thrombosis, hepatotoxicity liver failure and death[21]

2-2: PHILADELPHIA NEGATIVE NEOPLASMS

The term “neoplasm” was introduced in 2008 by the authors of the WHO Classification of Tumors of Hematopoietic and Lymphoid Tissues to underscore the clonal nature of myeloproliferative disorders.[12] Myeloproliferative Neoplasm (MPNs), are a group of rare hematologic malignancies that are characterized by excessive proliferation of hematopoietic precursors in the bone marrow leading to the abnormal production of one or more types of terminally differentiated blood cells such as platelets, white blood cells, and red blood cells. There are three main MPN diseases: Essential Thrombocythemia (ET), Polycythemia Vera (PV) and, Primary Myelofibrosis (PMF), which are also called Philadelphia negative MPNs to be distinguished from Chronic Myeloid Leukemia (CML). There are also two more progressed versions of MPNs, which are called POST-ET myelofibrosis and POST-PV myelofibrosis and are progressed versions of myelofibrosis originally developed from ET or PV. The excessive production of blood cells in all MPNs is a result of a single somatic mutation in HSCs. One single mutation can cause hyperplasia either in only one lineage or several lineages of hematopoiesis[32]. PV is characterized by erythrocytosis (increase in red blood cells), ET with thrombocytosis (increase in platelets)[33]. PMF, on the other hand, is more heterogeneous in terms of how different cell types change their population in blood, and it is mostly characterized by an excess of collagen fibers and megakaryocytic hyperplasia[32]. MPNs are very rare among

the world population; their incidence rate is lower than 6 per 100000 persons per year with PV from 0.4 to 2.8, ET from 0.38 to 1.7, and PMF from 0.1 to 1.0 per 100 000 persons per year[34]. These disorders generally occur in the middle- or advanced-age adults, with a median age of 65 years for PV, 68 years for ET, and 70 years for PMF[35]. MPN patients overall have reduced life expectancy compared with the general population; however, this lower survival rate in MPN patients is mainly related to death from hematologic malignancies or bacterial infections, and in young patients also from cardiovascular and cerebrovascular disease [36].

Here in this section, we tend to give an overview of Philadelphia negative MPN diseases in three main parts:

- The genetic basis of MPNs
- Risk scores and clinical diagnosis of MPNs
- Possible therapies of MPNs
- Disease progression

2-2-1: Genetic Basis

Since 2005, the discovery of genetic mutations has been a great help of understanding and management of MPNs. ET, PV, and PMF are stem-cell-derived; however, the complexity of clonal architecture and hierarchy in these diseases makes it difficult to predict the evolution of cell types populations[37]. So far, there are different mutations discovered to be reasons for MPN pathogenies. Among the most important and MPN-restricted mutations are JAK2V617F, JAK2 exon 12, MPL mutations, and CALR mutations.[38,39,40,41,42,43,44, 45]

JAK2V617Fmutation

The somatic JAK2 valine-to-phenylalanine (V617F) mutation has been detected in up to 90% of PV patients[46] and 50% to 60% of ET and PMFs.[47,48,49,39,50] Important note on JAK2V617 mutation is that it happens in a multipotent hematopoietic progenitor and, therefore, will be later detected in all myeloid cells and some of the lymphoid cells like B and NK and more rarely in T cellse[51]. V617F allele burden correlates with hematologic characteristics and clinical endpoints in MPN patients[52]. The highest JAK2V617 allelic burden is found in PV patients. ET patients are reported to have the lowest allelic burden, and most PMF patients display intermediate-high levels of allelic burden. V617F allele burden is close to 100%in post-PV or post-ETMF. [53, 54]

JAK2 Exon12

Another mutation in JAK and signal transducer and activator of transcription (STAT) pathway, which is found in V617-negative PV patients, is JAK2 exon 12 mutation. Those patients with a JAK2 exon12 mutation presented an isolated erythrocytosis and distinctive bone marrow morphology. Several of them also had reduced serum erythropoietin levels. [36]

Thrombopoietin receptor gene (MPL) mutations

Myeloproliferative leukemia virus mutations are another type of mutations that are found to be the root cause of MPN in some patients. MPL gene encodes for the TPO-R MPL, which signals through JAK2 and is considered essential for megakaryopoiesis[55]. The most frequent MPL mutation is called MPLW515L and K[43]. MPL mutations are reported in ET and PMF patients; however, the incidence rate is very low: 0% to 10% of PMF patients and in 0% to 6% of ET patients.[56]

MUTATIONS IN THE CALRETICULIN GENE: CALR

Late in 2013, another mutation, the CALR mutation, was discovered to play as a root cause of ET and MF,[45] and therefore, since then, it is accounted as a new molecular diagnostic marker. It is reported to be the second most common alteration after JAK2V617F in MPNs (50%-60%ET and 75%PMF) [57,58]. CALR protein plays essential roles, such as assisting many glycoproteins in folding, contributing to calcium homeostasis and participating in biological processes of proliferation, and apoptosis. [55] In *CALR*-mutant ET, the median allele burden reported to be around 49.5%, which is high in comparison to the JAKv617F mutation allele burden. *CALR*-mutant PMF has the median allele burden of around 52%.[59]

LYMPHOCYTE ADAPTOR PROTEIN LNK MUTATIONS

In 2010 other mutations were discovered to play roles as the drivers of the MPN disease. These mutations occur in LNK protein, which regulates JAK2 activation.[60]The lymphocyte adaptor protein (LNK) is a family member of adaptor proteins involved in cell signaling and control of B cell populations. It has a critical role in the regulation of signaling in hematopoiesis.[61]In most cases, LNK mutation is reported to be the secondary somatic mutation, which is associated with JAK2V617F or CALR mutations during disease progression.[62]

So far, the mentioned mutations are restricted only to MPN patients, which activate JAK2 kinase-dependent cytokine receptor pathways. There are three main mutations that are truly capable of driving MPN pathogenies. The current studies hypothesized that the phenotype of MPNs is mostly related to the types of activated receptors[32]. For example, JAK2 exon12 only causes an erythrocytosis phenotype. JAK2v617 is common among all three types of MPNs, so it causes phenotypes associated with ET, PV, and MF. CALR mutation does only happens in ET and MF and therefore resemble phenotypes of ET and MF. MPL activation plays a central role in changing the megakaryocytes (MKs) population. [43,63,64]

There are also non-restricted mutations that may be responsible for the heterogeneity of MPNs, which is not explained by the three main mutations. Most of these mutations are involved in phenotypic changes and disease progression in MPNs, and they are found to disturb proteins' activity in different cell functions such as DNA methylation regulator, histone modifiers, transcription factors, and splicing factors.[65,66,67]

Vainchenker and Kralovics [32]have summarized all the possible MPN mutations, their place of occurrence, and more information, such as their role in MPN pathogenies and their frequency nicely in their review on the genetic basis in MPNs. Table 3 is adapted from their study on the genetic basis in MPNs. Table 3 is adapted from their study.

2-2-2: MPN DIAGNOSIS AND RISK STRATIFICATION

In this section, we tend to give a review of the clinical and morphological features of the classical MPNs: ET, PV, and PMF.

2-2-2A: WHO CLASSIFICATION OF CLASSICAL MPNS

The WHO diagnostic criteria for classical MPNs have been revised in 2016. Table 4 is adapted from studies by Aber et al. [48] and Cazzola et al.[68] and summarizes the classification criteria for MPNs.

WHO classification is mostly based on the abnormal cell counts and bone marrow morphology changes; however, with the discovery of genetic mutations, the diagnostic criteria have been

updated and become more accurate. The diagnostic criteria are used in clinics for therapy management.

Three main cell types in peripheral blood whose abnormalities have a central role in diagnosis in MPNs are platelets, red blood cells, and white blood cells (WBCs). Thrombocytosis is defined as a platelet (PLT) count more than $450 \times 10^9/L$.[69] Although thrombocytosis may happen in all three MPNs, differential diagnosis of thrombocytosis is a major component of WHO classification criteria for ET diagnosis.[70]

An absolute erythrocytosis is present when there is an increase in the red cell mass over 125% of that predicted for the individual's body mass. [71]

Erythrocytosis can be found in all three MPNs, specifically in JAK2 (V617F)-mutant ET and PMF; however, erythrocytosis is a major diagnostic criterion for PV[72]. Because erythrocytes counts are reported either as hemoglobin level or hematocrit level, when Hemoglobin is higher than 16.5 g/L in men and higher than 16.0 g/L in women, the patients have the major criteria for being diagnosed as PV. Or in terms of hematocrit level, patients with hematocrit levels higher than 49% in men and higher than 48% meet the major criteria for PV diagnosis.

Another peripheral blood cell type is leukocytes or WBCs. An increase in leukocytes called leukocytosis. Leukocytosis is common in all MPN patients, but as leukocytes cover a great number of sub-cell types, in some MPNs such as PMFs there might be simultaneous increase and decrease of different white blood sub-cell types, therefore another diagnostic criterion may help for better and more accurate diagnosis.[73]

TABLE 3: LIST OF MPN MUTATIONS, GROUPED BASED ON THEIR FUNCTIONS

Signaling MPN driver	Other signalings	DNA methylation	Histone modifications	Transcriptions factors	RNA splicing
JAK2V617	LNK	TET2	EZH2	TP53	SRSF2
JAK2 EXON 12	CBL	DNMT3A	ASXL1	CUX1	SF3B1
MPL	NRAS	IDH1		IKLZF1	U2AF1
CALR	NF1	IDH2		ETV6	
	FLT3			RUNX1	

2-2-2B: BONE MARROW MORPHOLOGY

Bone marrow biopsy examination is vital in the diagnosis of MPN. From the bone marrow biopsy and changes in the morphology of different cell types, we have better insight to diagnose different MPNs. Megakaryocytes are large bone marrow cells with a large lobated nucleus. They are responsible for the production of blood thrombocytes (platelets), which are necessary for normal blood clotting. The number of megakaryocytes in ET is increased, and they are usually larger in size compared to normal megakaryocytes. Another feature of megakaryocyte in ET is that they are hyperlobulated, meaning they contain more than the usual number of lobules. The megakaryocytes also often form small clusters, which can be acknowledged on the biopsy specimen. [74] A picture of the ET bone marrow specimen is shown in figure 4A.

In PV patients, bone marrow shows a great hypercellularity, which is caused by prominent erythroid hyperplasia and scattered, increased megakaryocytes of varying sizes and shapes.[75] Hypercellularity of the bone marrow is shown in figure 4B. Different sizes and shapes of megakaryocyte are shown in figure 5.

PMF patients' hypercellularity of the bone marrow is due to an increased number of granulocytes. Megakaryocytes form clusters and show cytological atypia with abnormally lobulated, often hyperchromatic nuclei. The “Streaming” of cells is a sign of underlying marrow fibrosis. [76] Histology of bone marrow shown in figure 4C-E.

2-2-2C: GENETIC MUTATIONS IN MPNS AS A DIAGNOSIS CRITERION

We have done a short review of all the possible genetic mutation possibilities in the previous section (2-2-1). Genetic mutations are also a part of the diagnosis procedure based on WHO. The presence of JAK2 mutations, CALR, or MPL mutation is a major criterion for the diagnosis of ET. For ET diagnosis, it is also important that ET patients don't meet any of the WHO criteria for a for BCR-ABL positive CML, PV, PMF, myelodysplastic syndrome, or other myeloid neoplasms. As the high frequency of JAK2 mutations occurrence is reported in PV, the presence of JAK2V6717 or JAK2 exon12 mutation is a major criterion for PV diagnosis. For PMF patients, presence of JAK2, CALR, or MPL mutations or in the absence of these mutations, presence of another clonal such as (ASXL1, EZH2, TET2, IDH1/IDH2, SRSF2, SF3B1) is a major criterion for diagnosis.

TABLE 4: SUMMARY OF WHO CRITERIA FOR PHILADELPHIA-NEGATIVE MPNS

MPN type	Clinical and morphological features	Gene mutations	Complications
PV	-Erythrocytosis frequently combined with thrombocytosis and/or leukocytosis -Typically associated with suppressed endogenous erythropoietin production. -Bone marrow hypercellularity for age with trilineage growth	-JAK2 (V617F) in about 96% of patients -JAK2 exon 12 mutations in about 4% of patients	-increased risk of thrombosis -possible progression to myelofibrosis and less commonly to a blast phase similar to AML, -sometimes preceded by a myelodysplastic phase
ET	-Thrombocytosis. -Normocellular bone marrow with a proliferation of enlarged megakaryocytes	-JAK2 (V617F) in 60%-65% of patients - CALR exon 9 indels in 20%-25% of patients - MPL exon 10 mutations in about 4%-5% of patients - Noncanonical MPL mutations in, 1% of patients	-increased risk of thrombosis and bleeding, -possible progression to more aggressive myeloid neoplasms -JAK2 (V617F)-mutant ET involves a high risk of thrombosis and may progress to PV or myelofibrosis -CALR-mutant ET involves lower risk of thrombosis and a higher risk of progression to myelofibrosis

		-About 10% of patients do not carry any of the above somatic mutations (the so-called triple negative cases)	-Triple-negative ET is an indolent disease with a low incidence of vascular events
PMF	Prefibrotic PMF -Various abnormalities of peripheral blood -Granulocytic and megakaryocytic proliferation in the bone marrow with lack of reticulin fibrosis Overt PMF -Various abnormalities of peripheral blood. -Bone marrow megakaryocytic proliferation with atypia, accompanied by either reticulin and/or collagen fibrosis grades 2/3. -Abnormal stem cell trafficking with myeloid metaplasia (extramedullary hematopoiesis in the liver and/ or the spleen)	- JAK2 (V617F) in 60%-65% of patients -CALR exon 9 indels in 25%-30% of patients -MPL exon 10 mutations in about 4%-5% of patients -Noncanonical MPL mutations in, 1% of patients -About 5%-10% of patients do not carry any of the above somatic mutations (the so-called triple-negative cases)	- PMF is associated with the greatest symptom burden and the worst prognosis within MPNs, with a variable risk of progression to AML -CALR-mutant PMF is associated with longer survival compared with other genotypes - JAK2 (V617F)- and MPL-mutant PMF have a worse prognosis than CALR-mutant PMF -Triple-negative PMF is an aggressive myeloid neoplasm characterized by prominent myelodysplastic features and high risk of leukemic evolution

2-2-3: RISK STRATIFICATION

Although the WHO criteria for ET and PV diagnosis are different, PV and ET share very similar symptoms such as the high risk of thrombosis and hemorrhage or transformation to the secondary myelofibrosis or acute myeloid leukemia in both categories.[77]

There are prognosis scores invented to stratify the risk of thrombosis as a possible complication in MPNs. The following scoring systems are invented to estimate the risk of thrombosis in ET patients:

- The conventional score for prediction of vascular complications developed by European Leukemia Net[78].
- IPSET-thrombosis (International Prognostic Score for ET: estimates the risk of thrombosis)[79]
- IPSET (International Prognostic Score for ET: predicts survival) [80]

These systems categorize ET patients in two or three groups of developing thrombosis. In all these three scoring systems, age, platelet counts or leukocyte counts, and history of thrombosis or other cardiovascular risk factors, such as hypertension and diabetes, are the basic factors for risk stratification. Some of them additionally consider the JAK2 mutation. All

three scoring systems are brought in table 5, which is adapted from a paper review by Rumi and Cazzola [81].

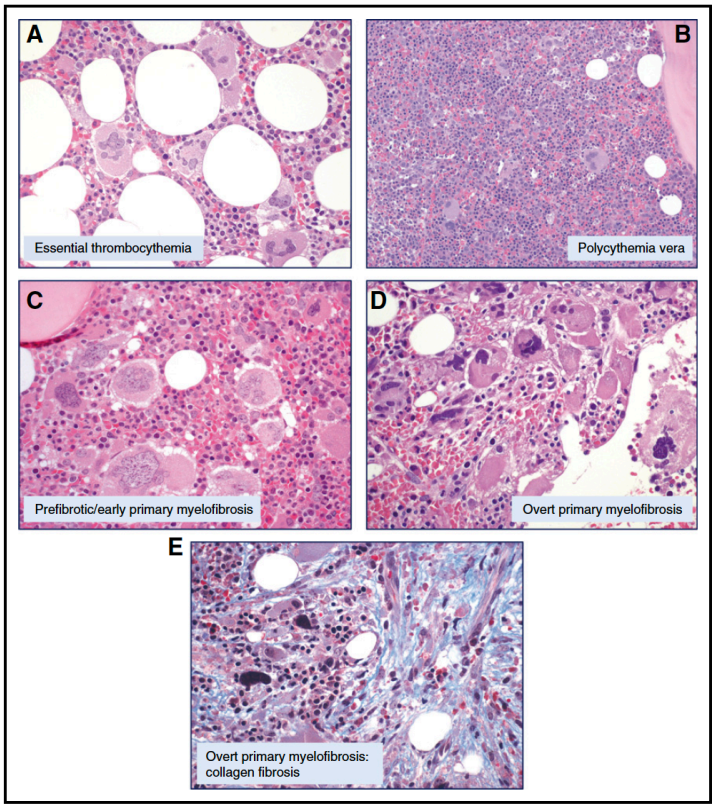


FIGURE 4: BONE MARROW HISTOLOGY OF DIFFERENT MPNS. THE FIGURE IS TAKEN FROM [72](COURTESY OF EMANUELA BOVERI, FONDAZIONE ISTITUTO DI RICOVERO E CURA A CARATTERE SCIENTIFICO POLICLINICO SAN MATTEO, PAVIA, ITALY.)

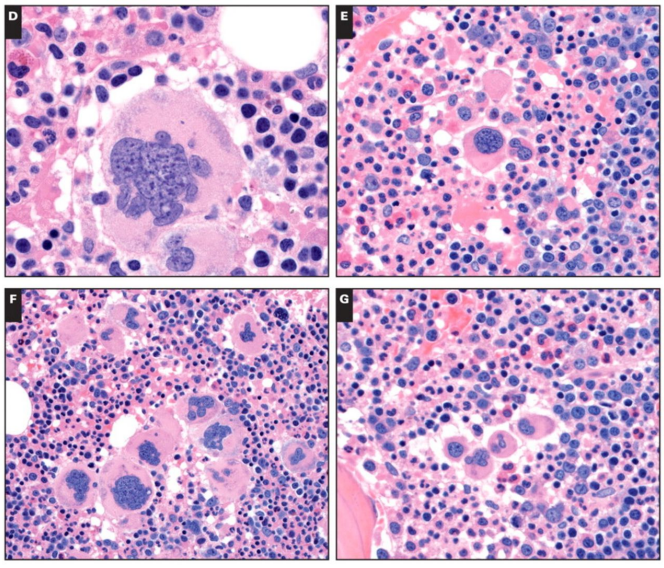


FIGURE 5: DIFFERENT CELL SIZES AND SHAPES OF MEGAKARYOCYTES IN PV. THE FIGURE IS TAKEN FROM [75].

In PV, there are two scoring systems as follows:

- The conventional score for prediction of vascular complications developed by European LeukemiaNet[78]
- IPSS for overall survival in PV[82]

Based on Leukemia Net, age and history on thrombosis are the basic risk factors for the development of thrombosis in PV patients. IPSS system considers the leukocyte counts additionally as a potential risk factor. The overview of prognostic models for thrombosis in PV patients is brought in table 6, which is adapted from a paper review by Rumi and Cazzola [81].

In PMF, there are three scoring systems as follows:

- IPSS International Prognostic Scoring System[83]
- DIPSS (Dynamic International Prognostic Scoring System[84]
- DIPSS-Plus [85]

In all of these systems, there are some risk factors whose presences as a symptom gets one point score, and the sum of scores can indicate the risk level. All of these three systems estimate the risk of thrombosis in PMF patients in four levels of low risk, intermediate-1, intermediate-2, and high risk. In IPSS and DIPSS age, hemoglobin level, WBC counts, circulating blast, and constitutional symptoms (such as weight loss, fever, headaches, fatigue, and…) are the risk factors with the difference that hemoglobin abnormalities have a higher weight to calculate the risk level in DIPPS. DIPSS-Plus also integrates unfavorable karyotype as a risk factor. The detailed information about how risk factors are used in calculating the risk level is brought in table 7, which is adapted from a paper review by Rumi and Cazzola[81].

2-2-4: TREATMENT IN MPNS

There are, in general, two main strategies to handle MPN patients, one is conventional treatments, which are mostly the use of cytoreductive drugs and the management of the symptoms of MPNs. The second strategy is the use of targeted therapies, such as JAK inhibitors. Since PV and ET have prolonged life expectancy, most of the conventional treatments in low-risk ET and PV patients have the goal of minimizing the risk of complications such as thrombosis and bleeding. Therefore, the possible therapies are aspirin, hydroxycarbamide (HC), interferon-alpha or anagrelide [77]and phlebotomy[86]. Treatment options for PMFs include hydroxyurea as a cytoreductive agent to attenuate hyperproliferative manifestations such as leukocytosis and thrombocytosis[87]. Possible treatments for anemia, which is one of the most frequent manifestations in PMF patients with inferior survival, are the use of androgens, erythropoiesis-simulating agents, immunomodulating drugs, chronic transfusions and iron chelation (removal of excess iron from the body with special drugs)[88]. Splenomegaly is another symptom in PMF patients, which is defined as the enlargement of the spleen. There are possible pharmacological treatments, like interferon-α and in case of unsuccess, splenectomy, and splenic irradiation are possible therapies[88]. The only curative approach for PMF patients so far is stem cell transplantation [77]. With the discovery of JAK2 mutation in MPNs in 2011 and later on the other mutations, targeted therapies such as ruxolitinib are invented to treat myelofibrosis (MF) and are being developed as the therapy option for hydroxyurea resistant or intolerant PV and ET[89]. The two studies COMFORT-I /II evaluated the ruxolitinib and showed spleen volume reduction and improvement in symptoms in MF patients[90,91]. RESPONSE, RESPONSE2, and RELIEF are three studies that evaluated the effect of ruxolitinib in HC intolerance PV patients and showed effective controls of blood count and spleen size, symptoms and in some cases reduction in thrombotic events[92,93,94,95].

Verstovsek and colleagues[96] tested the effectiveness of ruxolitinib in ET patients and showed that ruxolitinib might be helpful in the control of blood counts, spleen size reduction, and other symptoms.

There are also other JAK inhibitors such as fedratinib, pacritinib, and momelitinib, which were stopped to be used further due to toxification or worsening the symptoms[97,98,99,100,101,102]. There have also been some studies to examine the effect of the combination of JAK inhibitors and traditional cytoreductive drugs; however, the results are not validated[77].

TABLE 5: SCORING SYSTEMS FOR THROMBOSIS RISK ASSESSMENT IN ET PATIENTS

<table>
<tr><td colspan="3">The conventional score for prediction of vascular complications developed by European Leukemia Net</td></tr>
<tr><td rowspan="3">Risk factors (at least one of the risk factors)</td><td>age≥ 60</td><td>Risk groups</td></tr>
<tr><td>Previous thrombosis or major bleeding</td><td rowspan="2">Low risk: age <60 y AND no history of thrombosis or major bleeding AND PLT count <1500 × 109/L, that is, none of the 3 major risk factors

High risk: age ≥60 y AND/OR history of thrombosis or major bleeding AND/OR PLT count ≥1500 × 109/L, that is, at least 1 of the 3 major risk factors</td></tr>
<tr><td>Platelet count ≥ $1500 \times 10^9/L$</td></tr>
<tr><td colspan="3">IPSET-thrombosis (International Prognostic Score for ET: estimates the risk of thrombosis)</td></tr>
<tr><td rowspan="4">Risk factors (each risk factor has a weight. Overall risk is the sum of points)</td><td>age≥ 60 (one point)</td><td>Risk groups</td></tr>
<tr><td>Previous thrombosis (two points)</td><td rowspan="3">Low risk: 0-1 point (probability of thrombotic events: 1.03% of patients/year)

Intermediate risk: 2 points (2.35% of patients/year)

High risk: ≥3 points (3.56% of patients/year)</td></tr>
<tr><td>Cardiovascular risk factors, such as hypertension, diabetes, and active tobacco use (one point)</td></tr>
<tr><td>JAK2v617 mutation (two points)</td></tr>
<tr><td colspan="3">IPSET (International Prognostic Score for ET: predicts survival)</td></tr>
<tr><td rowspan="3">Risk factors (each risk factor has a weight. Overall risk is the sum of points)</td><td>age≥ 60 (two points)</td><td>Risk groups</td></tr>
<tr><td>Previous thrombosis (one point)</td><td>Low risk: 0 (median survival not reached)</td></tr>
</table>

	Leukocytes count $\geq 11 \times 10^9/L$ (one point)	Intermediate risk: 1-2 points (median survival, 24.5 years) High risk: 3-4 points (median survival, 13.8 years)

There are also emerging targeted therapies other than JAK inhibitors such as drugs acting on the PI3K/Akt/mTOR pathway, drugs acting as farnesyltransferase inhibitors, histone deacetylase inhibitors (HDACi), telomerase inhibitors, hedgehog pathway inhibitors, hypomethylating agents, hedgehog pathway inhibitors, angiogenesis inhibitors, antifibrotic agents[77].

TABLE 6: SCORING SYSTEMS FOR THROMBOSIS RISK ASSESSMENT IN PV PATIENTS

The conventional score for prediction of vascular complications developed by European Leukemia Net		
Risk factors (at least one of the risk factors)	age≥ 60	**Risk groups**
	Previous thrombosis or major bleeding	Low risk: age <60 y AND no history of thrombosis that is no risk factors High risk: age ≥60 y AND/OR history of thrombosis that is at least one risk factor
IPSS for overall survival in PV		
Risk factors (each risk factor has a weight. Overall risk is the sum of points)	age≥ 67 (5 points)	**Risk groups**
	$57 \leq age \leq 66$ (2 points)	Low risk: 0 (median survival of 28 years) High risk: ≥3 points (median survival of 11 years)
	Previous thrombosis (one point)	
	Leukocytes count $\geq 15 \times 10^9/L$ (one point)	
	Previous thrombosis (one point)	
	Leukocytes count $\geq 11 \times 10^9/L$ (one point)	

2-2-5: DISEASE PROGRESSION IN MPNS

Since the main aim of our study is modeling disease progression in MPNs, in this section, we reviewed studies on disease progression in MPNs and search for state of the art in this field.

All three forms of classic MPNs share several clinical characteristics, such as clonal hematopoiesis, bone marrow hypercellularity, and developing complications such as thrombosis and hemorrhage. There are also some differences such as elevated erythrocyte counts in PV, elevated platelet levels in ET, and bone marrow fibrosis in PMF, which have elaborated previously[103]. In general, MPNs are chronic diseases with prolonged life expectancy, but some patients may progress to more evolved stages. Progression stages include post-PV or post-ET secondary myelofibrosis (SMF), which are associated with an increase in reticulin fibrosis in the bone marrow and extramedullary hematopoiesis[104]. There is also another type of progression, called "accelerated phase" (AP), which is characterized by the deficiency of all three cellular components of the blood (red cells, white cells, and platelets) and the blast counts up to 20% in the bone marrow[105]. There are some studies reporting rare cases of MPN progression to AML as a major complication with a life expectancy of 5 months. A notable risk factor for such complications is blast count exceeding than 20 %[106].

Risk factors for leukemic transformation in all MPNs are brought in table 8, which is adapted from a study by Yogarajah and Tefferi [107].

TABLE 7: SCORING SYSTEMS FOR THROMBOSIS RISK ASSESSMENT IN PMF PATIENTS

IPSS		
Risk factors (each risk factor has a weight. Overall risk is the sum of points)	age≥ 65 (one point)	**Risk groups**
	Constitutional symptoms (one point)	Low risk: 0 (median survival of 11.3 years) Intermediate-1 risk: 1 point (median survival of 7.9 years) Intermediate-2 risk: 2 points (median survival of 4 years) High risk: ≥ 3 points (median survival of 2.3 years)
	Haemoglobin $< 10g/dL$ (one point)	
	WBC count $> 25 \times 10^9/L$ (one point)	
	Circulating blasts $\geq 1\%$ (one point)	
DIPSS		
Risk factors (each risk factor has a weight. Overall risk is the sum of points)	age≥ 65 (1point)	**Risk groups**
	Constitutional symptoms (one point)	Low risk: 0 (median survival of at least 20 years) Intermediate-1 risk: 1-2 points (median survival of 14.2 years)
	Haemoglobin $< 10g/dL$ (two point)	

	WBC count $> 25 \times 10^9/L$ (one point)	Intermediate-2 risk: 3-4points (median survival of 4 years)
	Circulating blasts $\geq 1\%$ (one point)	High risk: 5-6 points (median survival of 1.5 years)
DIPSS-plus		
Risk factors (each risk factor has a weight. Overall risk is the sum of points)	DIPSS score (DIPSS low=0, DIPSS-int1=1 point, DIPSS-int2= 2 points, DIPSS-high = 3 points)	**Risk groups**
	Red blood cell transfusion need (one point)	Low risk: 0 (median survival of 15 years) Intermediate-1 risk: 1 point (median survival of 6.6 years)
	Platelet count $<100 \times 10^9/L$ (one point)	Intermediate-2 risk: 2-3 points (median survival of 2.9 years)
	Unfavorable karyotypes (one point)	High risk: 4-6 points (median survival of 1.3 years)

Lundberg and colleagues[108] have concluded in their research that the total number of somatic mutations in MPNs was inversely correlated with survival and risk of leukemic transformation. They developed a model which is relating the somatic mutations with the progression of the disease. Figure 6, which is taken from their paper, show their model in detail. Based on this model, depending on the oncogene types and the number of mutations, MPN patients may transform sooner or later to the leukemic phase, or they may show superior or inferior survival time. As depicted in their model, about 10% of MPNs have unknown driver mutations. This group of MPNs has a lower risk of disease progression. About 54% of MPNs have either JAK2 mutation or CALR mutation. These group also show the less progressive type of MPNs and in the spectrum of disease progression are placed after the first group. When there are more additional somatic mutations to CALR or JAK2 mutation, the MPNs are closer to transforming to AML in the progression spectrum. In another study [103], Klampfl and colleagues investigated the role of chromosome abnormalities in the chronic phase and during the disease progression. They suggested that an increased number of chromosomal lesions was significantly associated with patient's age, as well as with the disease progression and leukemic transformation. They found abnormalities of some chromosomes to be positively associated with the progression to secondary myelofibrosis and abnormalities in some other chromosomes to be associated with leukemic transformation.

Another study[109] supports the idea that the JAK2 V617F positive chronic myeloproliferative disorders show a biological continuum with phenotypic presentation partly influenced by JAK2 V617F mutational load. Other studies [110,111] suggest that chronic inflammation is either a promotor of mutagenesis in MPNs, or it may imply an increased risk for the development of other cancers. Most of the research in modeling disease progression in MPNs either are in the field of understanding genetic mutations and developing clonal markers, or they study the relationships between clinical biomarkers and survival of MPNs. In the field of mathematical modeling of MPN disease progression however; unfortunately, there has not been enough

research and a need for developing a computational method that is able to track disease progression is extremely felt. In the next section, we review existing computational models for disease progression in MPNs and CML.

TABLE 8: RISK FACTORS FOR LEUKEMIC TRANSFORMATION IN MPN PATIENTS

Risk factors	ET	PV	MF
Clinical	age≥ 60 thrombosis	age≥ 70 or age≥ 61	age≥ 65 red blood cell transfusion dependency
Laboratory	Platelet count >1000 × $10^9/L$ ANEMIA Leukocytes count ≥ 5 × $10^9/L$	Leukocytes count ≥ 15 × $10^9/L$	Platelet count <50 × $10^9/L$ Leukocytes count ≥ 30 × $10^9/L$ PERIPHERAL BLAS ≥3% and/or Platelet count <100 × $10^9/L$ Interleukin 8, and interleukin 2 receptor C-reactive protein > 7mg/L and peripheral blast >1%
Bone marrow	Prefibrotic primary myelofibrosis morphology, reticulin grading, and bone marrow cellularity		Bone marrow blasts ≥10%
Previous therapy		Cytoreductive agents, pipobroman, P32, chlorambucil	-Abnormal karyotype -Chromosome 17 aberrations -Monosomal karyotype and/or complex -Unfavorable karyotype complex or sole
Cytogenetic abnormalities	None	Abnormal karyotype	
Mutations	TP53, EZH2	SRSF2, IDH2	Triple-negative mutational status ASXL1, SRSF2, IDH1, or IDH2

2-3: INTRODUCTION ON MATHEMATICAL/COMPUTATIONAL MODELING IN THE FIELD OF MPNS

Cancer is a genetic disease initiated by a mutation and evolved through the accumulation of more and more mutations leading to heterogeneity and uncontrolled growth of tumor cells and finally, a breakdown of system[112,113]. Although cancers are inherently heterogeneous, there are some general functional rules, often called *hallmarks of cancer*. These rules include the ability of cancer cells to stop their apoptosis and block immune response, which results in the evolution of cancer and ultimately paves the way to metastasizing[114,115]. After somatic damage, a normal cell turns into a tumor cell that is able to evade apoptosis. Accumulation of more somatic changes leads to the exponential growth of transformed cells. Therefore, we have two main level transformations: first cell-level, which is responsible for cancer initiation and second population level, which is responsible for cancer progression[116]. At the population levels, heterogeneous tumor cells population dominate over healthy differential cells population[117].

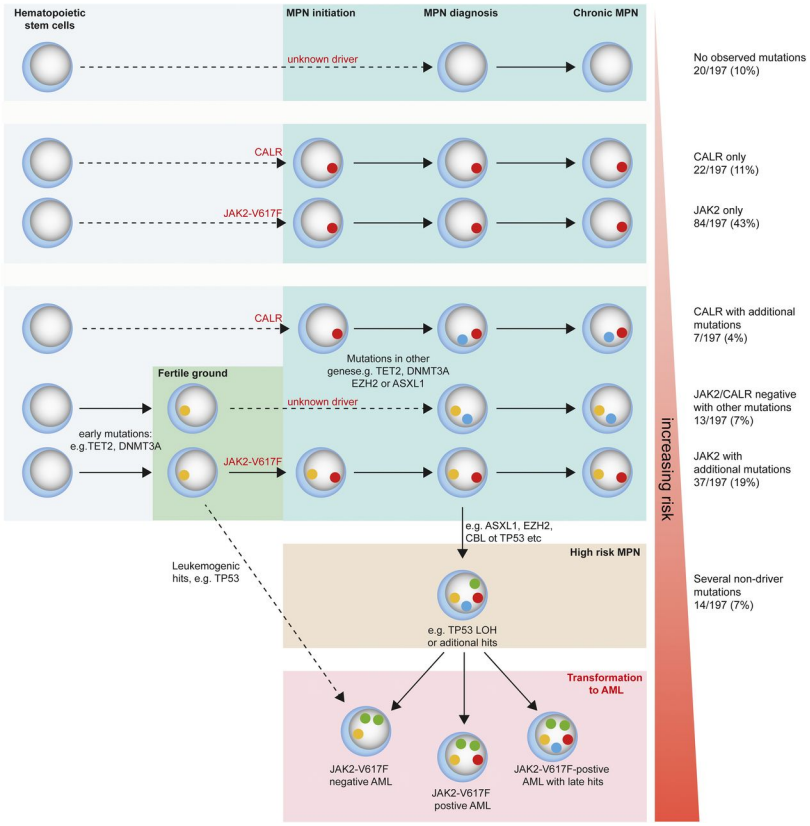


FIGURE 6: TAKEN FROM [108] PRESENTS A MODEL OF MPN DISEASE EVOLUTION AND RISK STRATIFICATION IN CORRELATION TO MUTATIONAL EVENTS,

Recently Computational cancer modeling has taken a great interest among researchers[118]. There are two big arms in the computational modeling of cancers. One is called mechanistic modeling, which develops a hypothesis about cancer growth and then translates this hypothesis in mathematical formulations to provide quantitative insight on cancer formation/evolution and help for a better understanding of underlying biological mechanisms. In mechanistic modeling, there have been developed models that concentrate on modeling cancer initiation and progression by looking into biochemistry and biology of cancer[119, 120,121]. Usually for developing mechanistic models, there should be a good understanding of biological

mechanisms that drive cancer. However, as they are based on prior knowledge about the system, they are not data-dependent, and they can be extrapolated on data except the one used for parameter fitting. There is also some research focusing on the biophysics of cancer, claiming that simulating the thermodynamics of a system can help to predict cancer phase transition to more progressed states[117]. In the other arm, researchers seek help from machine learning algorithms, such as regression models, support vector machines classifiers, and neural networks to model cancer evolution out of data. These methods, in general, are called data-driven modeling. As data-driven models are fully data-dependent, one of their big problems is their inability to be extrapolated on new data. Machine learning techniques and artificial intelligence have opened a new area in cancer prediction, as they provide easier integration of data from heterogeneous data sources such as gene expression data, next-generation sequencing (NGS) data, and clinical lab data. To overcome the problems with mechanistic and data-driven modeling, if possible, one can integrate both mathematical and data-driven models to set up more robust models for predictions. These methods are called hybrid models and used previously in the chemical industry [122]. As the focus of our study is modeling disease progression in MPNs and CML, we review the existing models for these types of blood cancers.

2-3-1: CML MODELING

2-3-1A: MECHANISTIC MODELS: POPULATION DYNAMICS MODELS

The first mathematical modeling of CML studies goes to back to 1991 when Fokas and colleagues developed a mathematical model for granulocytopoeisis[123]. Their model deals with seven different blood cell types, which play a role in the process of granulocytopoeisis. They include stem cells, three proliferative cell types, myeloblast, promyelocyte, myelocyte, and three non-proliferative cell types. They set up a mathematical formulation based on the number of times each cell type proliferates to the next lineage and the time each cell type spends in each cell lineage.

In 2004, Moore and colleagues modeled the interaction between naive T cells, effector T cells, and CML cancer cells, with a system of three ordinary differential equations (ODEs) describing the population change rate of these three cell types [124].

Michor and colleagues in 2005 developed a mathematical model based on the data from imatinib-treated patients.[119] Their model includes in overall three major cell categories, including healthy cells, CML cells without resistant mutations, and finally cells with imatinib-resistant mutations. Each major cell category consists of the stem cells, progenitor cells, differentiated cells, and terminally differentiated cells leading to overall 12 ODEs. Each ODE describes the population change rate of the mentioned cell types. This model is a basic ground for further development in the field of modeling dynamics of CML, such as a model developed by Dingli and colleagues in 2008[125]. Dingli model integrates more than four cell types. Their model integrates a 32-compartmental hierarchy of hematopoiesis[126]. In 2005, Adimy and colleagues described the evolution of the cell population based on a system of 2 non-linear differential equations with a distributed time delay corresponding to the cell cycle duration. Then by constructing a Lyapunov functional, they showed that trivial equilibrium is globally asymptotically stable if it is the only equilibrium and a nontrivial equilibrium becomes unstable via a Hopf bifurcation, which is the most biologically meaningful situation [127].

In 2008 Wodarz, introduced a mathematical model[128] to support his theory that, a transition from chronic phase to blast crisis does not need the invoke of further mutations and can be

explained solely by feedback mechanisms regulating the patterns of stem cell division, in particular, the occurrence of symmetric versus asymmetric cell division. In 2012, Zhang and colleagues[129] integrated both Moore's[124] and Michor's models[119] to characterize the dynamics of CML infection and the development of the mutations. In 2014, Flå and colleagues[130] used the existing models to deduce the bifurcation patterns in the normal and leukemic stem cells' populations in CML. In 2015, Ledzewics and Moore[131] introduced a mathematical model treatment in CML. Specific to their model is considering immune system effects with one extra compartment and separation of the CML cells into quiescent and proliferating classes. In 2017, Woywod and colleagues[132] developed a group of deterministic mathematical models to study the origins of clinically observed dynamics of CML. Each model is a system of coupled first-order differential equations, considering populations of three mutated active stem cell strains and three associated pools of differentiated cells; two models allow for activation of quiescent stem cells.

2-3-1B: DATA-DRIVEN MODELS

Another field of computational modeling in CML research is data-driven modeling and the use of machine learning methods. Here we summarized some studies which applied artificial intelligence methods for a better understanding of CML.

Sasaki and colleagues[133] used neural networks on the response data of 630 CP- CML patients to predict the cumulative incidence of major molecular response with one year of TKI treatment. Khosla and Ramesh [134]applied convolutional neural networks on images of blood samples of CML patients in CP, AP, and BC stages of CML and developed an image classifier that enables the classification of different phases of CML. Dincer and colleagues[135] presented a deep profile framework that extracts latent variables from a big set of gene expression data sets of leukemia-related disease, including CML, by using variational autoencoders to predict the phenotypes. In another study on microscopic chromosome image data, Wang and colleagues, [136] developed automated identifiers of abnormal metaphase chromosome cells for the detection of CML. Yeung and colleagues[137] integrated gene expression data and expert knowledge by using iterative Bayesian model to identify a small set of genes that are highly predictive of CML phases and used the selected genes to predict relapse after allogeneic transplantation in CML patients. Oehler and colleagues[138] used the bayesian model averaging algorithm to a large set of DNA arrays of CML patients and defined a small set of genes capable of identifying CML progression from the chronic phase to blast crisis.

2-3-2: MPN MODELING

In contrast to CML, which is intensively studied with the help of mathematical modeling and probably is the most studied cancer because of its simple nature, mathematical modeling for MPN is a new field, and there is a great potential for more research and developing progression models. In the following, there is a short overview of a few studies on the computational modeling of MPNs.

In 2009 Haeno and colleagues[3] designed a stochastic mathematical model to describe the evolution of mutations in a differentiation hierarchy of hematopoiesis. Their goal was to answer ambiguities about the origin of myeloid malignancies. The motivation of their model was questioning this hypothesis that the cancer progression constitutively arises always from stem cells. Therefore, they tested this assumption: In the special case of JAK2v617F positive MPNs, if two mutations are required for MPNs, the first mutation, which is not JAK2v617, occurs in *committed progenitors* to confer properties of self-renewal and longevity required for the

secondary mutations and the development of phenotypic MPNs. Their model considers different evolutionary pathways for cancer-initiating cells in JAK2 positive MPNs:

- If JAK2V617F mutation occurred in a *stem cell*.
- If firstly a mutation (other than JAK2v617F) occurred in *a progenitor cell* inducing self-renewal ability, and JAK2V617F mutation happened secondly.
- If JAK2V617F mutation happened firstly in a *progenitor cell*, and a mutation happened secondly to induce the self-renewal ability.
- If firstly, a self-renewal inducer mutation occurred in stem cells, without changing stem cell phenotypes and JAK2V617F mutation occurred secondly in a progenitor cell.

The result of their mathematical model confirms that JAK2V617 happens most likely in a progenitor cell.

In 2010 Attolini and colleagues[139] developed a computational approach called RESIC, Retracting the Evolutionary Steps in Cancer, to investigate the temporal sequence of somatic genetic events in cancer. They initially developed their approach on colorectal cancer data sets, but then later, they validated their approach on a set of secondary AML samples with preceding MPN and known JAK2/TET2mutations. They concluded that for these MPN patients, JAK2 most likely proceeded TET2.

In the most recent study in the field of mathematical modeling of MPNs, Andersen and colleagues[140] suggested a population dynamics model for JAK2v617 positive patients as a proof of concept that chronic inflammation is a trigger of clonal evolution and disease progression in MPNs [110,111]. Their model is an extended version of a previously developed mathematical model for cancer cells[141] by coupling inflammatory feedback as a regulator of the mutational rate, self-renewal rates of both healthy and cancerous cells, and apoptosis.

CHAPTER 3: METHODS AND MATERIAL:

In this chapter, we try to provide the main strategies we used in modeling MPNs, including the applied concepts and machine learning algorithms. There is also a description of the materials used in this study.

3-1HYBRID MODELING

In the area of systems modeling, to which systems biomedicine also belongs, there are, in general, two main strategies of modeling: Data-driven modeling and mechanistic modeling. The main goal in systems modeling is to model the behavior of a system either by understanding the interfering mechanisms or by investigating the relations between input and output data, which are results of experiments in different conditions. In mechanistic modeling, also called white-box modeling, mechanisms are fully understood and can be represented in terms of mathematical equations. Examples of use cases mechanistic model use cases are modeling cell signaling, metabolism, population dynamics in disease modeling, organ modeling, etc. Straightforward interpretation, extrapolability, and low data demand are some of the advantages of mechanistic modeling. For establishing a reliable mechanistic model, however, there should be sufficient knowledge and understanding about involving mechanisms. In systems biomedicine, since we are studying complex living organisms, most of the time the access to this knowledge is restricted. Therefore, data-driven modeling can be more helpful. In data-driven modeling, there is no access to a full understanding of involving mechanisms, and the goal is to extract as much as possible information about the relation between input and output data. Development of biomarkers for diagnosis and prognosis, genome-wide association studies, and therapy planning are some application areas of data-driven or black-box modeling. Black-box models are models that can approximate any input-output function of a system by small errors; however, to simulate the system with a small error, we should have access to a large amount of data. One disadvantage of black-box models is, as its name suggests, the suggested model is blind with respect to the involving mechanisms of the system. Black-box models also have this disadvantage that they fail to be extrapolated on a dataset outside the convex hull of the training dataset. Extrapolating from a black-box model can result in nonsensical prediction, which indeed is problematic in the field of precision medicine where accurate predictions are vital.

Due to the lack of sufficient knowledge about interacting disease-driving mechanisms, developing entirely mechanistic models for disease progression is very unlikely[1]. On the other hand, data-driven models cannot provide any insight with respect to disease-driving mechanisms. Hybrid modeling, also named grey-box modeling, is a combination of mechanistic and data-driven modeling. It is an integration of smaller sub-models based on different knowledge sources and can be either data-driven or mechanistic models. Hybrid models, therefore, can combine the benefits of data-driven models and mechanistic models. A Hybrid model consists of two general types of sub models[122]:

a) mechanistic sub-models for parts of the system, which the relation between input and output data is well understood
b) black box sub-models for parts of the system, which there are no rigorous model is available

By partitioning a big black box into smaller black boxes, the number of inputs feeding new smaller sub-models reduces. Reduced number of inputs results in less complexity, which

arises as a result of the curse of dimensionality[122]. Also, Fiedler and colleagues[142] showed that the extrapolative simulations outside the database could be possible by hybrid models.

As already mentioned in the introduction, in this study, we focus on MPNs. We have discussed different MPNs from their genetic basis to symptoms in chapter 2, and here it is only sufficient to recall that MPNs are a group of blood disorders including chronic myeloid leukemia, polycythemia vera, essential thrombocythemia, and primary myelofibrosis. However, in a more general classification, MPNs are divided into Philadelphia positive (CML) and Philadelphia negative MPNs (ET, PV, and PMF)[12]. Based on this classification of MPNs, we considered two different strategies for modeling: Hybrid modeling for CML and hybrid modeling for Philadelphia negative MPNs. Hybrid Modeling of CML takes the benefit of available population-dynamic-based mechanistic models together with the systems theory as a tool for the detection of critical transition points in the progression of CML. As there is no available mechanistic model for the evolution of Philadelphia-negative MPNs, we used data-driven and hybrid methods for training smaller sub-models on divisions of feature space as a strategy for better classification of Philadelphia negative MPNs. (figure 7)

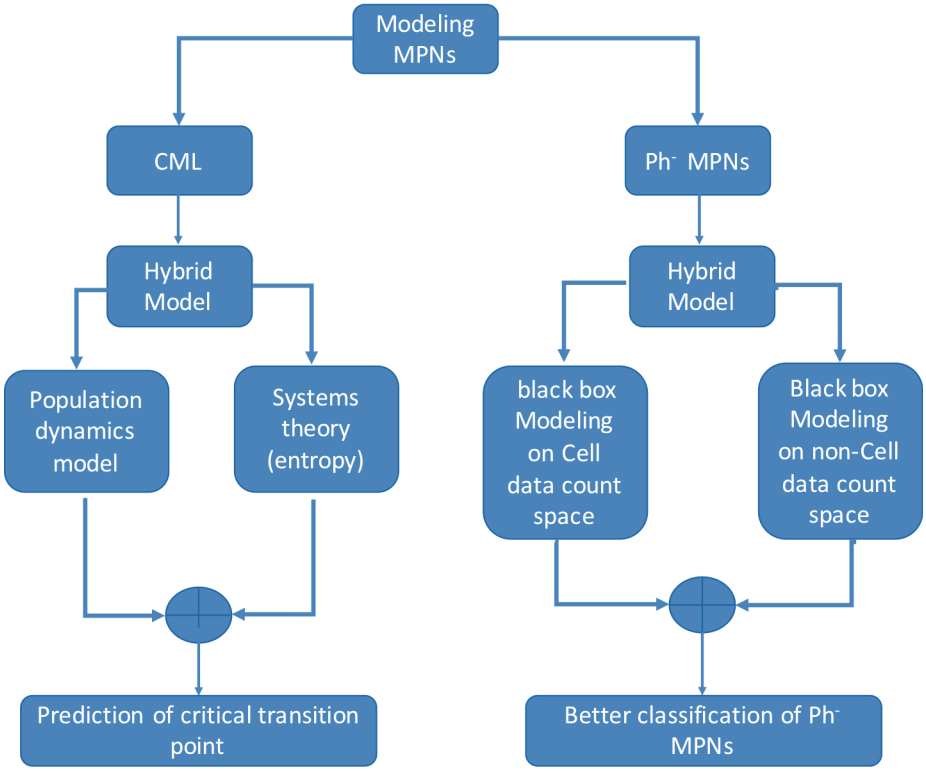


FIGURE 7: THE WORKFLOW OF THE PROJECT. DEVELOPMENT OF HYBRID MODELS FOR CML AND MPNS IN DIFFERENT STRATEGIES.

Therefore, in this chapter, we try to describe concepts such as systems theory and machine learning algorithms used for developing hybrid models. We also provided conceptual descriptions of some algorithms such as Principal Component Analysis (PCA) and PhysioSpace, which are used as a preprocessing method before applying machine learning algorithms.

3-1-2 Phase Transition:

Biological systems are complex and often exhibit non-linear behavior. With this point of view, diseases such as cancer can be seen as a result of perturbations. These perturbations can be

studied by identifying sudden shifts in the state of the system. Identification of these sudden shifts is helpful for better clinical management and therapy planning. These sudden shifts are called critical transition points, where a small perturbation causes the system states to shift from one to another[5]. Change in the dynamics of the system may serve as early warning signals for critical transitions [143]. Therefore, we should look for an observable variable with a considerable change in its dynamic around the transition point. There are different variables available for quantification and characterization of chaotic systems; however, some of them are sensitive to some parameters such as initial conditions, or they may experience rapid divergence in their evolution. Entropy is a physical quantity that is insensitive to the start point of the system[144]. Entropy is a thermodynamic concept that is used for quantification of disorder in a system and can serve as a warning signal for detecting critical transition points. In an equilibrium state, the entropy is at its lowest; however, at a critical transition point, the entropy is at its maximum [145]. We used entropy as a biomarker for the identification of transition point in chronic CML (chapter 4).

3-1-3 MACHINE LEARNING ALGORITHMS

Before jumping into the conceptual description of machine learning algorithms that we used in this study, it is proper to mention a general grouping of the machine learning methods. In a general view, we have two categories of machine learning: supervised and unsupervised.

In supervised learning, we provide the desired algorithm, the pairs of input and output data, and the algorithm task is to find the mapping function between input and output. It is called supervised since we know the output for each input. So, the algorithm can correct itself. Supervised learning tasks include classification and regression. In a classification task, the output is a finite set of values such as low- high or benign-malignant, or stage one-stage two-stage three. In a regression task, the output value is a real value such as room temperature, and weights.

In contrast to supervised learning, we provide only input values to the unsupervised algorithm, and the algorithm finds natural patterns inside the input data space regardless of output. In contrast to supervised learning, there are no output data provided to the algorithm, and the algorithm itself should discover interesting patterns inside the data. Clustering, dimension reduction, and finding associations are some of the unsupervised learning problems.

As we will see in chapter 5, there are three black boxes in the structure of our MPN hybrid model (figure 42). The first and third black boxes are MPN classifiers. The second black box is a regression model for the prediction of misclassification errors of the first black box. We used neural networks, logistic regression, support vector machine, and random forest as machine learning algorithms to train these three black boxes.

To compare our hybrid model with other modeling strategies, we used K-means and hierarchical clustering. Principal component analysis and PhysioSpace are also unsupervised methods that we used for gene expression analysis of MPNs. In the following, we provided a description of the algorithms used in our research.

3-1-3-1 SUPERVISED LEARNING ALGORITHMS

3-1-3-1-A NEURAL NETWORKS

Neural Networks (NN) are machine learning algorithms inspired by the model of the human brain and are entirely different from conventional digital computers[146]. In analogy with the

human brain, NNs consist of an interconnected network of simple processing units called perceptron or neuron. A NN has two operative modes: the training mode and the action mode[147]. In the training mode, a neuron learns to solve a problem from experience and examples (data) by modifying its connections[148]. In this mode, they are trained to detect a pattern in input data. In action mode, by detection of the trained input pattern, neurons produce output data[147]. A neuron in detail is shown in figure 8. Each node receives a collection of signals from the previous layer and based on an activation function, generates an output which is fed as input to the neurons of the next layer. Each connection between a pair of nodes has a weight, w_i. The output of a neuron, y_i, is calculated as follows:

$$y_i = f(b + \sum W_i \times x_i)$$

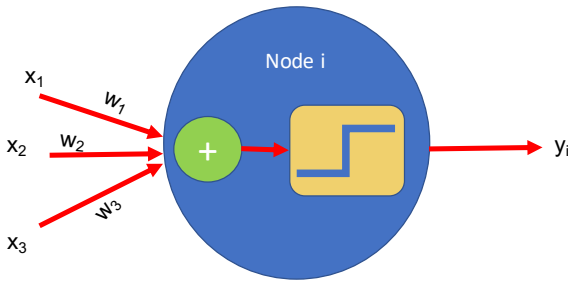


FIGURE 8: DETAILS OF A NEURON

The activation function determines if the summation of received signals are enough to fire a neuron. It can be different functions such as the step function, the sigmoid function, or the tanh function. When an NN receives input data, which are, for example, the values of independent variables in a dataset, it tries to learn a pattern. Pattern recognition is done by minimization of the output errors each time a new pattern (input data point) is fed into the network. There are different structures of NN, such as feed-forward NN and feedback or recurrent NN. A feed-forward NN is shown in figure 9. In an overall view, it consists of multiple layers. The signal flow is one-sided and only from input to output. Different layers are:

- Input layer
- Hidden layers (can be one or multiple layers)
- Output layer

In a feed-forward NN, there is no loop from outputs to inputs, which means the output of each layer is effective only on the next layer. In a feedback NN, signal flow is in both directions, and outputs of each layer can actively also affect the current layer.[149]

Training an NN on an available dataset means to calculate the weight of every single edge in the whole network. Backpropagation, short for "backward propagation of errors," is an algorithm for supervised learning of artificial networks using gradient descent. Given an artificial neural network and an error function, the method calculates the gradient of the error function with respect to the neural network's weights.

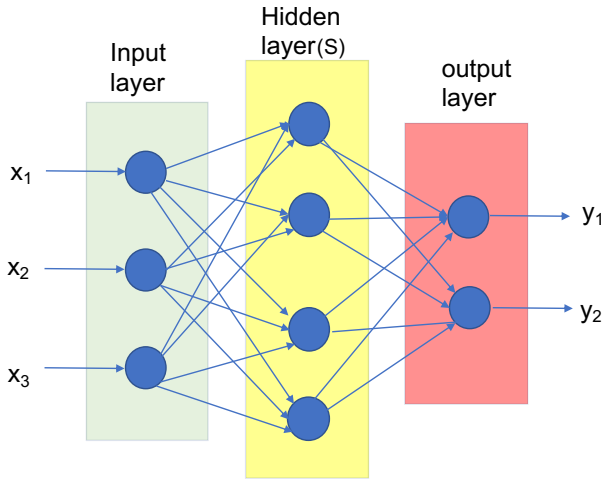


FIGURE 9: STRUCTURE OF A FORWARD NEURAL NETWORK

In figure 9, each blue circle represents a neuron, and each arrow is a connection between a pair of neurons.

NNs are achieving more and more interest because of their high applicability in different areas of data-driven modeling. Some of their advantages over the conventional techniques, according to Jack v. Tu[150] are:

- They are not limited by the prior assumption of normality, linearity, and variable independence.
- They can detect non-linear relationships between independent and dependent variables.
- They can detect all possible interactions between predictors.
- They can be developed using multiple different training algorithms.

Another advantage of NNs is that they can be easily implemented in parallel architectures, which results in a drastic reduction of the processing time compared to other kinds of algorithms, achieving similar results. Another priority of NNs is that due to their parallel nature when an element of the NNs fails, other parts continue their function without any interruption. Despite their ability to solve modeling problems, NNs also have some disadvantages. As NNs are black-box models, it is tough to interpret the causal relationships between the input and output data. They often need a huge dataset for proper learning. Besides, artificial NNs have some inherent disadvantages such as slow convergence speed in training, less generalizing performance, arriving at a local minimum and over-fitting problems.

Neural networks are used vastly in data-driven modeling. Their applications, however, most fall into major categories of classification, prediction, and optimization[151]. They can be used in different areas such as computer vision, financial analysis, medical research for diagnostic, prognostic, and prediction purposes[152].

3-1-3-1B LOGISTIC REGRESSION

Logistic regression is one of the machine learning techniques which is used mostly for the binary classification problems. Logistic regression can answer questions such as what is the probability of mortality in diabetic patients (yes or now) if they are obese and smoking. Alternatively, classifying breast cancer patients into two different stages based on their gene expression profile. Logistic regression is based on a logistic function which maps a set of input variables to predict binary dependent variables. The logistic function is defined as:

$$f(x) = \frac{1}{1 + e^{-(\propto x + \beta)}}$$

Based on this function, we can map all values in $[-\infty, +\infty]$ into [0,1] interval. Based on a probability value that is calculated from logistic function and compared with a threshold value (which is often 0.5), we can classify our data point either as class one or class two [153].

When using logistic regression for classification problems, we are looking for proper parameters $\propto$ and β in the logistic function formula. We do that by optimizing a cost function, which is usually maximum likelihood function [154].

Logistic regression separates input data space into „regions" by a linear boundary. Therefore, to use logistic regression it is required that data is linearly separable (Figure 10).

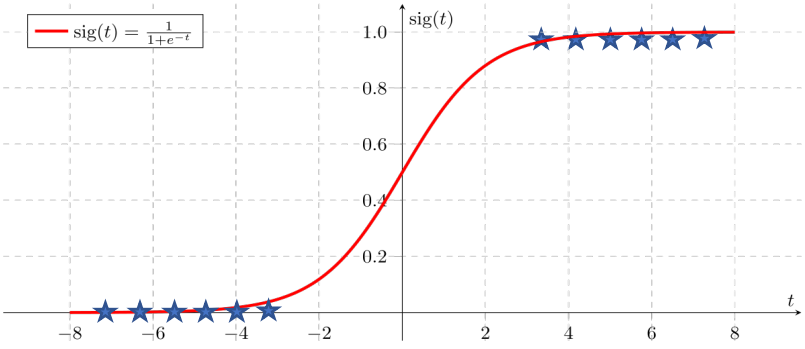


FIGURE 10: LOGISTIC FUNCTION. A LOGISTIC FUNCTION MAPS THE NEGATIVE VALUE OF T (LOWER STARS) TO 0 AND A POSITIVE VALUE OF T (UPPER STARS) TO ONE.

Although logistic regression is used basically for binary classification, we can extend its applicability also to multinomial classification problems. For this purpose, we should apply some tricks to convert a multiple-classes problem to multiple binary classification problems. One versus all and one versus one classification techniques are some of these tricks. In both solutions, instead of one classifier, we have to train multiple classifiers. The advantages of logistic regression are its simplicity, high efficiency, and high interpretability. It can be easily regularized and is convenient to be implemented. These advantages make it an excellent option to start with. One disadvantage of logistic regression is that we cannot solve nonlinear problems with logistic regression, and its application is limited to classification i.e., it is not useful when the output has a continuous value.

3-1-3-1C Support Vector Machines

Support vector machines (SVMs) are machine learning algorithms that are used vastly in classification tasks. SVM learns from training the data to predict the labels for objects [155]. Unlike the logistic regression, which the linearity of data space is an assumption of proper use, SVM can be applied not only for linear but also for nonlinear applications (figure 11).

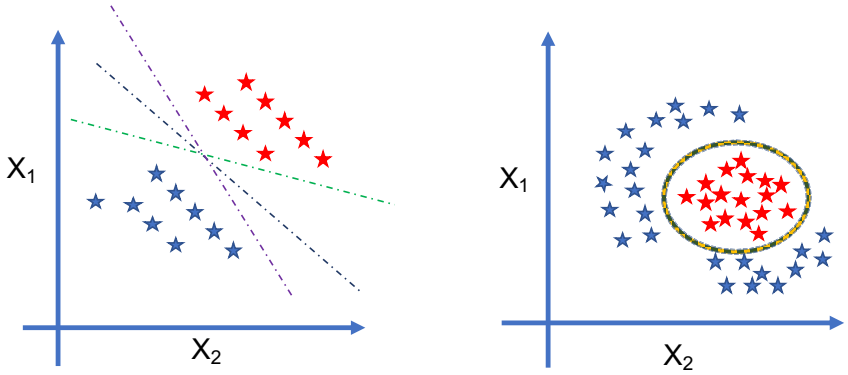


FIGURE 11: SVM IS APPLICABLE BOTH WHEN FEATURE SPACE IS LINEARLY SEPARABLE (LEFT) OR NON-LINEARLY SEPARABLE

To understand how SVM works, there are four basics that must be learned[156]:

- the separating hyperplane
- the maximum-margin hyperplane
- support vectors
- the kernel functions

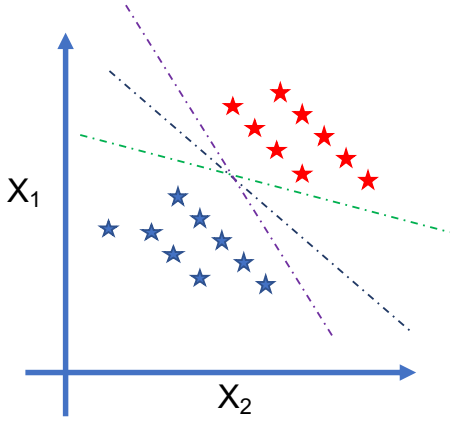


FIGURE 12: SEPARATING HYPERPLANE SHOWN AS DOTTED LINES.

In figure 12, there are different lines to separate blue stars as class 1 from red stars as class 2. These lines are called *separating hyperplanes*. In this case of 2D data, hyperplanes are lines, but for data with more dimensions, we need surfaces and hyperplanes. SVM tries to select the best separating hyperplane. To find this optimal solution, SVM *maximizes the margin* around the separating hyperplanes. Margin, in this definition, is the distance between the hyperplane and the closest data points to hyperplane (margin is shown as the green box in figure 13. *Support vectors* are the training data points that in case of their removal, the position of separating hyperplane would change. In fact, support vectors are the input vectors that touch the boundary of margin (stars inside circles on the boundary of margin are support vectors in figure 13).

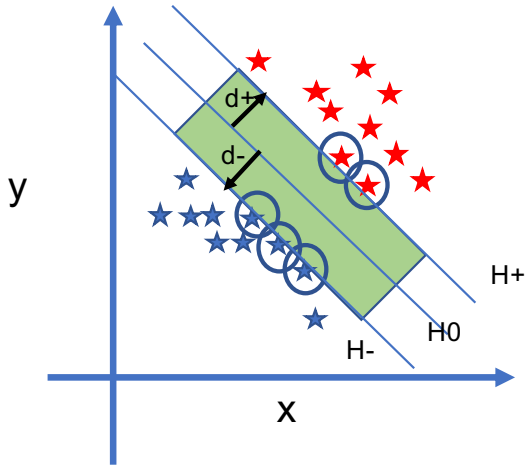


FIGURE 13: MARGIN IS SHOWN AS A GREEN BOX, STARS IN THE CIRCLES ARE CALLED SUPPORT VECTORS.

Let us define class +1 for the red stars and class -1 for the blue stars in figure 13. We look for the optimal hyperplane (line H0 in figure 13) such that:

$$\begin{cases} w \times X_i + b \geq +1 \; when \; y_i = +1 \\ w \times X_i + b \leq -1 \; when \; y_i = -1 \end{cases}$$

H+ and H- in are the planes:

$$\begin{cases} H+ : w \times X_i + b = +1 \\ H- : w \times X_i + b = -1 \end{cases}$$

The separating hyperplane H0 is where $w \times X_i + b = 0$. SVM tries to find the H0 by maximizing the margin.

So far, we talked about SVM on a linearly separable data space. However, for example, in figure 11-right, we see that a linear hyperplane fails to perfectly separate blue and red stars, and using a linear hyperplane will result in very low accuracy. Instead, a circular hyperplane (yellow-green circle in the figure 11-right) can perfectly separate the red stars from the blue stars. *Kernel functions* are transformation functions that convert non-linear feature space into linear feature space. There are different types of kernel functions, such as polynomial kernels, Gaussian kernels, radial basis functions, etc.

Like other classification techniques, there are some advantages and disadvantages of using SVMs. SVMs apply to a wide range of classification problems like problems in high dimensions and non-linearly separable feature spaces. In contrast to NNs, which because of the iterative nature of their learning processes, such as backpropagation, suffer from converging only to locally optimal solutions, SVMs always find a global minimum. Also, the geometric interpretation of SVMs provides fertile ground for further investigation. One disadvantage of SVM is that the selection of a right class of kernel is necessary for a good classifier and as the selection of the right kernel is a discrete optimization problem, SVMs are slower machinery for large datasets in comparison to NNs.

3-1-3-D Random Forests

Random forest (RF) is another data mining technique that is used both for classification and regression problems. As the name suggests, random forests are a collection of tree-structured

classifiers called decision trees [157]. A decision tree is a partitioning method that divides the feature space into smaller subspaces based on output value[158]. Decision trees are useful for several applications such as classification and prediction, variable selection, assessing the relative importance of variables, etc. [159]

To understand how a decision tree works, I provided a simple example. Suppose that we have a binary output Z (0 or 1) for a 2D input space (x, y) (figure 14-left). Z=0 is shown as red stars and Z=1 as blue stars. In this example, attribute (feature) space has two dimensions: x and y. The first step is to look for the attribute causing minimal partitions; here, in this case, it is better to start firstly with dividing x value. The best threshold is x=1.5. For x>1.5 subspace, we have both blue and red stars, which can be separated if we consider a horizontal line at y=1.3. For y >1.3 we have only red stars which are equivalent to Z=1 and for y<1.3 we have only blue stars i.e., Z=0. In x<1.5 subspace, we need more than two partitions. The first partition is achieved when we consider a horizontal line y=1. For y<1 we have only red stars, but for y>1 we have both red and blue stars, which can be separated from each other by considering another vertical line at x=0.9. For x>0.9 we have only red stars and for x<0.9 we have only blue stars. The decision tree for this example is shown in figure 14-right. A decision tree, as depicted in figure 14-right, has nodes and edges. Nodes are the attributes of the variable space, and edges are the partitions on the corresponding node. Last level nodes are in fact, the output.

Decision trees are the basic building block of RFs. An RF is a collection of multiple random decision trees. Each decision tree provides a prediction value for the asked problem. The final prediction provided by RF is an aggregation of all the decision trees' predictions. RF has shown excellent performance in datasets with more variables than observation[160].

Another important concept of building RF is a bootstrap aggregation or bagging. Bootstrapping is done for the improvement of stability and accuracy of an algorithm. Suppose we have a dataset D with a size of n (n= number of data points). In bootstrapping, we sample randomly and with replacement n' data points out of n. Replacement here means that one data point can be selected multiple times. n' is usually ~63.2% of n [161]. We can repeat the sampling for T times. As a result, in the end, we have T random subsets of data. On each random subset, we train an algorithm. In the case of random forests, this algorithm is a decision tree. In the end, we have T decision trees. Each decision tree provides a prediction for a new sample. The final result for the new sample is an average of T predictions in the case of regression or voting in the case of classification.

The inventors of random forest, Bierman and Cutler[162], believe that RFs are fast to train and suitable for large data sets. It can handle high dimensional input space without variable deletion and estimate which variables play an essential role in classification. RF is also useful for estimating the missing value and is one of the most accurate models.

Although decision trees are easy to interpret, as RFs are ensembles of a significant number of decision trees, they are hard to interpret. Although Bierman and Cutler claim that RF never overfit, but in case of noisy data, like other machine learning algorithms they may encounter overfitting.

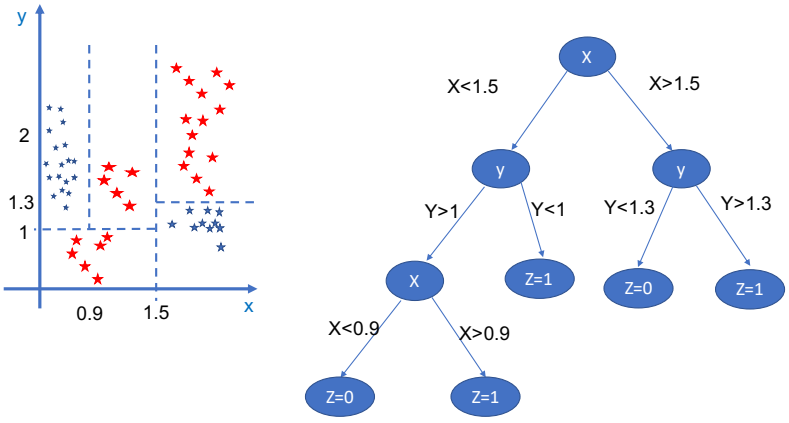


FIGURE 14: AN EXAMPLE OF HOW A DECISION TREE WORKS. LEFT: DATA DISTRIBUTION IN A 2D FEATURE SPACE. COLOR OF STARS SHOWS THEIR OUTPUT VALUE. RIGHT: DECISION TREE FOR DATA DISTRIBUTION ON LEFT

3-1-3-E CROSS-VALIDATION

As we already described some of the popular machine learning algorithms, there are different options and models that can be used on a dataset for further predictions. Cross-validation (CV) is often used for evaluating models and selection of the most proper model for a data set[163]. The basic idea of cross-validation is dividing the whole data set into two parts of training and test. The training parts are used for training an algorithm, and the test part is used for estimating the risk of the trained algorithm [164]. Based on CV results, we can pick an algorithm with the highest accuracy or lowest risk. There are different types of CV. The most popular CVs are train-test split and k-fold cross-validation. In the train-test split CV, the data is randomly split into normally 70:30 or 80:20 fractions. An algorithm is trained on the bigger set and evaluated on the smaller test. In K-fold CV, a dataset is randomly split in K folds. Then by leaving out the Kth fold, a candidate algorithm is trained on K-1 folds and tested on Kth fold. This procedure is being repeated until every K-fold is used as the test set. The performance metric is an average of all K repetitions. CV is used only as a tool for model selection. For example, we used CV in our research to pick between logistic regression, SVM, NN, and RF for a better description of our data.

3-1-3-2 UNSUPERVISED LEARNING ALGORITHMS

3-1-3-2-A PRINCIPLE COMPONENT ANALYSIS

Principal component analysis (PCA) is a statistical procedure invented by Karl Pearson as an orthogonal transformation of a set of correlated variables into a set of linearly uncorrelated variables, called principal components (PC)[165]. PCA is used as a tool for dimension reduction in high dimensional data. PCA is usually done based on the eigenvalue decomposition of the data correlation matrix [166]. PCA usually is performed prior to modeling algorithms such as linear regressions, neural networks, etc. To reduce data dimensionality and complexity of models.

In the following, I describe the process of PCA in figure 15.

Suppose that our data has two dimensions of X_1 and X_2. In figure 15, it can be seen that X_1 and X_2 are linearly correlated. The goal of PCA in this example is to reduce dimensions of data from 2 to 1. First, we shift the center of the coordinate system to the center of the data. Then we look for a direction with the highest variation. The first PC is the dimension aligned with the

highest variation and is achieved by rotation of the coordinate system. The rotation angle (θ) should be calculated such that new axes of the new coordinate system have a correlation of zero (because they should be orthogonal).

In a more general example, suppose that we have an $m \times n$ matrix with m as the number of observations and n as the number of variables.

The workflow of calculating principal components of this matrix are as follows:

- Normalizing the data by subtracting the mean from each of the data dimensions, which is equivalent to shifting center of the coordinate system to the center of data:

$$X_{new} = X - < X >$$

- Calculating covariance matrix C
- Calculate the eigenvalues (λ_k) and eigenvectors (u_k) of C

$$Cu_k = \lambda_k u_k$$

Eigenvectors u_k are mutationally orthogonal and axes of the new coordinate system. λ_k are the independent variances of the data distribution.

- Choosing a set of components and forming a feature vector

In this step, the data is compressed by picking the eigenvectors with the highest eigenvalues as new dimensions. The highest eigenvalues correspond to the highest variance in data. The selected set of eigenvectors is called the feature vectors.

- Deriving the new dataset

The final step in PCA is to calculate values on the new coordinate system composed of selected feature vector as follow:

$$final\ data = Row\ Feature\ Vector \times Row\ Data\ Adjust$$

Where *Row Feature vector* is transposed feature vector *and Row Data Adjust* is the mean-adjusted data transpose.

PCA has some limitations. To run PCA, some assumptions need to be true. These assumptions are the linearity in observed data, statistical importance of mean and covariance, and the association of significant variances with important dynamics in data. PCA leads to sub-optimal results when there is non-linearity in data or when there is an outlier. There are, however, alternatives for PCA such as Factor Analysis and Non-Negative Factor analysis, tSNE, and diffusion maps.

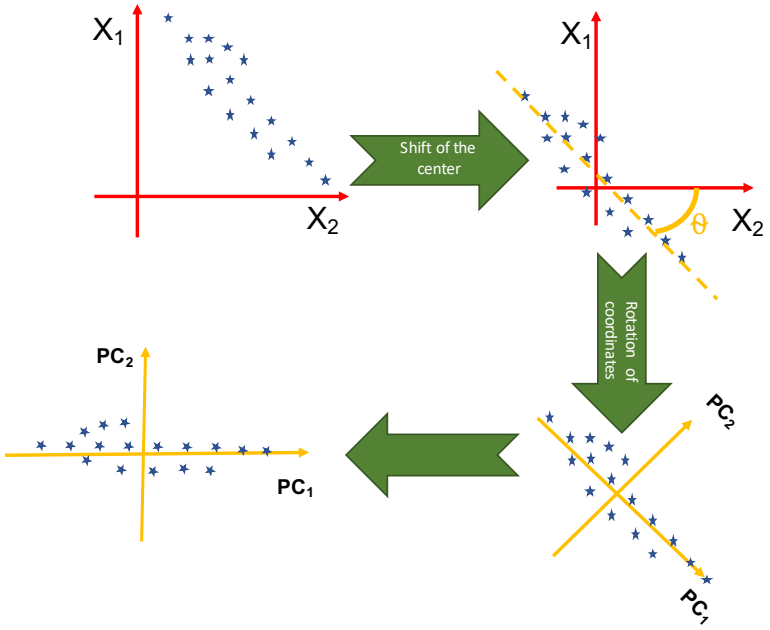


FIGURE 15: AN EXAMPLE OF PCA STEPS. 1) SHIFT TO THE CENTER OF DATA. 2) ROTATING THE COORDINATE SYSTEM SUCH THAT THE DIRECTION WITH THE HIGHEST VARIANCE BECOMES THE FIRST DIMENSION OF THE NEW COORDINATE SYSTEM. 3) DERIVING NEW DATASET

3-1-3-2-B PHYSIOSPACE

PhysioSpace is a method introduced by Lenz and colleagues[167], which relates gene expression experiments from heterogeneous sources using shared physiological processes. We used PhysioSpace as an unsupervised method for finding an association of MPN patients' gene expression profiles with different cell types or blood-related diseases. The goal of PhysioSpace, in general, is to compare the similarities between two physiological processes, such that similarity score is robust with respect to scaling and different data protocols.

Let us consider two processes P1 and P2 as:

P1: physiological state A => physiological state B

P2: physiological state C => physiological state D

We can define a similarity score, so-called PhysioScore, which can be calculated in three steps:

Step 1: calculating the distance /difference between the two processes.

Calculate the differential expressions for all genes x between states A (x_A) and B (x_B):

$$P_{AB} =: \quad p(A=>B) = -\log10(p(x_A,x_B))*sign(<x_B>_B - <x_A>_A)$$

We do the same for the second process (P2):

$$P_{CD} =: \quad P(C=>D) = -\log10(p(x_C,x_D))*sign(<x_C>_C - <x_D>_D)$$

The p-values $p(x_C,x_D)$ and $p(x_A,x_B)$ can be calculated using a t-test.

Step 2: the selection of most significantly expressed genes:

We select the 5% most upregulated genes in P_{AB}, and call them *u*. We select the 5% most downregulated genes in P_{AB} and call them *d*.

Step 3: Mapping of Process P1 on Process P2:

We calculate the physio score of p_{phys} (P1=>P2) which means how similar P1 is to P2 by the following formula:

$$P_{phys} = -\log10(p(P_{CD}(u), P_{CD}(d))) \times sign(<P_{CD}(u)>- <P_{CD}(d)>)$$

In step 3, the p-values $p(P_{CD}(u), P_{CD}(d))$ must be calculated using *Wilcoxon-Ranksum test*, as we are interested the how up/downregulated genes are sorted. A *t-test* is not sufficient.

3-1-3-2-C K-MEANS CLUSTERING

K-means algorithm is an unsupervised machine learning algorithm that tries to find inherent groupings inside a dataset. Each data point belongs to only one group based on a similarity measure. Each cluster is defined with its centroid. The basic idea is that the data points inside one cluster are as much as possible, similar to each other, and distinct to other subgroups. In other words, a data point belongs to a particular cluster if it is closer to that cluster's centroid than any other centroid. The critical steps of the K-means algorithm are brought as followings:

1. The first step is *a priori* to select the number of clusters, here in this example, suppose we decide for k=2
2. Select randomly two (because of k=2) data points as centroids of the clusters.
3. Calculate the distance between each pair of data points and the selected centroid.
4. Each data point belongs to a particular cluster if it is closer to that cluster's centroid than another cluster's centroid.
5. Update the clusters' centroids by averaging (arithmetic mean) over all the points inside each cluster
6. Iterate steps 3 to 5 until the centroids are fixed and unchanged.

K-means clustering is usually the first option that comes to mind to detect inherent subgroups inside a dataset. However, as the clusters are defined as a centroid, it is proper to use K-means, when the clusters have normal distributions or when the distance between clusters is large enough. When, for example, variance in one dimension (variable) significantly dominates other dimensions and the between-cluster distance is not large enough, K-means clustering fails in detecting the right clusters. In these cases, it is better to use other clustering techniques such as K-nearest neighbors.

3-1-3-2-D HIERARCHICAL CLUSTERING

Another clustering method that we used in this study is hierarchical clustering. The hierarchical clustering algorithm starts by considering each data point as a separate cluster. It successively mergers clusters together based on a similarity measure until a stopping point. The similarity measures can be, for example, Euclidian distance. In larger distances, clusters are more separable. The key steps in hierarchical clustering are:

1. Consider each observation as its own cluster.
2. Calculate the distance between all possible data point pairs. If we consider data points as vectors of size m (number of observations), the result of this step is calculating a so-called distance matrix of size $m \times m$.

3. Based on distance values, the closest pair is merged to make a cluster. So, the number of clusters reduces to $m-1$.
4. Having a new number of clusters, a new distance matrix is calculated with the size of $m-1 \times m-1$.
5. Again, the closest pair is merged to make a cluster. So, the number of clusters reduce to $m-2$. Steps 3 to 4 are being repeated until all clusters merged to one single cluster.

Hierarchical clustering is often shown in the dendrogram, which is a tree-like structure. Suppose our data consist of 6 points: x1, x2, x3, x4, x5, x6. A dendrogram as the result of hierarchical clustering for these data points is shown in figure 16:

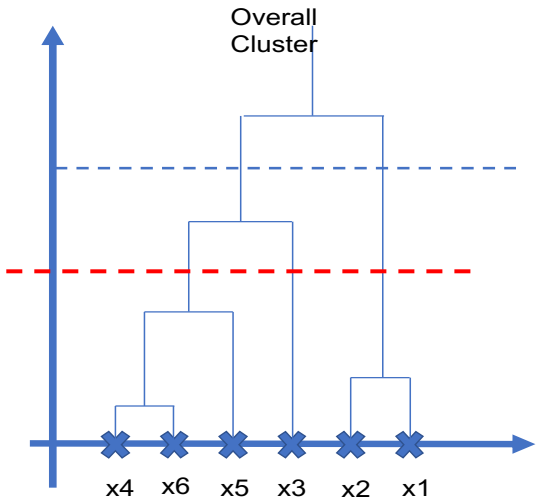


FIGURE 16: DENDROGRAM OF A HIERARCHICAL CLUSTERING

In figure 16, y-axis is the distance calculated between pairs. Firstly, each data point is a cluster of its own. In the next step, since the distance between x4 and x6 is less than others, x4 and x6 merge to make a new cluster. x1 and x2 also merge to form another cluster. Successively x5 and x3 join the first cluster. In the end, we have one overall cluster. If we want to have two clusters, we can cut the dendrogram with a horizontal line (blue dash line in figure 16). As a result, x1 and x2 are in one cluster, and the rest belongs to another cluster. If we want to have three clusters, we can cut the dendrogram with another horizontal line (red dash line). In this case, we have x1 and x2 in one cluster, x3 as a cluster of its own, and x4, x5, and x6 as another cluster.

Hierarchical clustering is easy to implement, and because of providing a structured dendrogram, they give information about the data. However, for large datasets, it is a slow clustering algorithm and may not be time-efficient.

3-2 MATERIALS

In this section, we describe the data that we used in hybrid modeling for CML and Philadelphia negative MPNs.

3-2-A MATERIALS FOR CML MODELING

In our hybrid model for CML, as it is brought in chapter 4 in details, we integrated two heterogeneous data sources:

- Simulated data from a mechanistic model which describes population dynamics of blood cells in chronic CML
- patients' gene expression profiles

Simulated data is achieved by implementing a model developed by Dingli[125]. The result of the implementation is a matrix of size 64×1336. 64 is the number of healthy and cancerous cell types in CP-CML and 1336 is, in fact, the sampled time points of the continuous evolution of the cell population.

For the gene expression analysis, we used GSE4170[168]. GSE4170 dataset contains the gene expressions of bone marrow cell mixtures of 42 samples in CP-CML, 9 samples in AP-CML, 8 samples in AP_{cyto}-CML, and 6 samples of healthy CD34+ cells.

3-2-B MATERIALS FOR PHILADELPHIA NEGATIVE MPN MODELING

As we will see later in chapter 5, Modeling Philadelphia negative MPNs is done in two separate sections based on the type of dataset:

- Modeling based on gene expression data
- Developing a hybrid model based on clinical data

For the gene expression analysis, we used two data sets:

The first data set is GSE26049[169] . The array for this datasets is an Affymetrix Human Genome U133 plus2.0. It contains purified RNA from the whole blood tissue, amplified to biotin-labeled aRNA, and hybridized to microarray chips. There are 21 healthy control samples, 19 ET patient samples, 41 PV patient samples, and 9 PMF patient samples. The number of genes in this dataset is 17275.

The second dataset, GSE120362[186], is from a study done by Czech and colleagues from Uniklinik Aachen. It contains expression profiling by array of human CD34+ cells of either bone marrow or peripheral blood of 32 MPN patients. The array type they used is Gene ChipT Human Transcriptome Array 2.0. There are 4 Healthy control samples, 6ET samples, 10 PV samples, 9 PMF samples, and 3 secondary myelofibrosis samples. It contains 67528 probe Ids.

And for hybrid modeling at the clinical data level, we used data from the German Study Group in MPNs (GSG-MPN)[170] with 572 patients in 5 different MPN stages: ET, PV, PMF, Post ET-MF and Post PV-MF and with more than 200 variables, including cell count data, age, sex, therapy information and etc.

CHAPTER 4: DEVELOPING A MODEL FOR TRACKING DISEASE PROGRESSION IN CML

As discussed in the introduction on CML, so far, there are prognostic scores such as Sokal[17], Hasford[18], and EUTOS[19] developed for risk stratification and survival prediction in CML patients. These scores are only based on established clinical parameters, such as blood cell counts, number of blasts, and spleens size. On the other hand, the computational models that are so far developed for CML mainly focus on describing the dynamics of the cell populations and explain how a single mutation in HSC leads to an abundance of white blood cells in peripheral blood. They are not generally predictive of progression to AP and BC. All in all, we believe that there is still potential to invent a biomarker for individualized risk assessment. There are studies that report genes specific to each state of CML [168], but as they provide a long list of genes, they are unfortunately not sufficiently reliable for clinical use [138]. However, there are studies that show the remarkable stability of genome-wide differential expression-based biomarkers [171,172].

Here we focus on developing a mathematical model that integrates the existing dynamic cell population models with gene expression-based biomarkers for a better risk assessment in CP-CML patients. The goal of our research is, that with the help of dynamics of the cell population in the chronic phase to invent a biomarker capable of inferring patient's status in the time course of disease evolution from their genomic profile. The goal of our study and the workflow of our method is shown in figure 17. The novelty of our approach is integrating different sources of data, such as the simulated cell population profile of different hematopoietic cell types and patients' genomic data.

4-1: MOTIVATION

The idea of combining population dynamics models and gene expression data in cancer modeling comes from the assumption that the observed changes in gene expression of a mixture of cells are in fact, the results of the cell population shifts. To support this assumption, we have compared the gene expression profile of a mixture of cells (GSE4170[168]) and gene expression profile of different cell types (GSE47927[173]) in different stages of CML. GSE47927 contains gene expression data obtained from four CD34+ enriched and flow-sorted CML stem and progenitor cell subpopulations (HSC: hematopoietic stem cell, MEP: megakaryocyte-erythroid progenitor, GMP: granulocyte-macrophage progenitor, CMP: common myeloid progenitor) across all phases of CML progression (CP, AP, APcyto: accelerated phase (additional clonal cytogenetic changes without and BC). This dataset contains GEX of 12 patients: six CP-CML, four AP-CML, and two BC-CML, and three healthy volunteers.[173] GSE4170[168] dataset contains gene expression of bone marrow cell mixtures of 42 samples in CP-CML, 9 samples in AP-CML, 8 samples in AP_{cyto} -CML, and 6 samples of healthy CD34+ cells.

To compare these two datasets, we did PCA [165] to overcome the high dimensionality of gene expression arrays. However, before applying PCA directly on the full genome arrays, as a preprocessing step, we extracted a subset of the 6,384 genes. These genes are previously shown[171] to be characterized as information-rich by a low information ratio (IR), corresponding to predictive accuracy with respect to clinical phenotypes and stable biomarkers[171]. Not all of these 6384 genes are commonly presented in the mentioned datasets due to different array technologies. Therefore, the number of common genes between low IR and available genes

in arrays, reduced to 6,078 genes for dataset GSE47927 and 4,309 genes for GSE4170. After preprocessing, we applied the PCA routine of MATLAB[174].

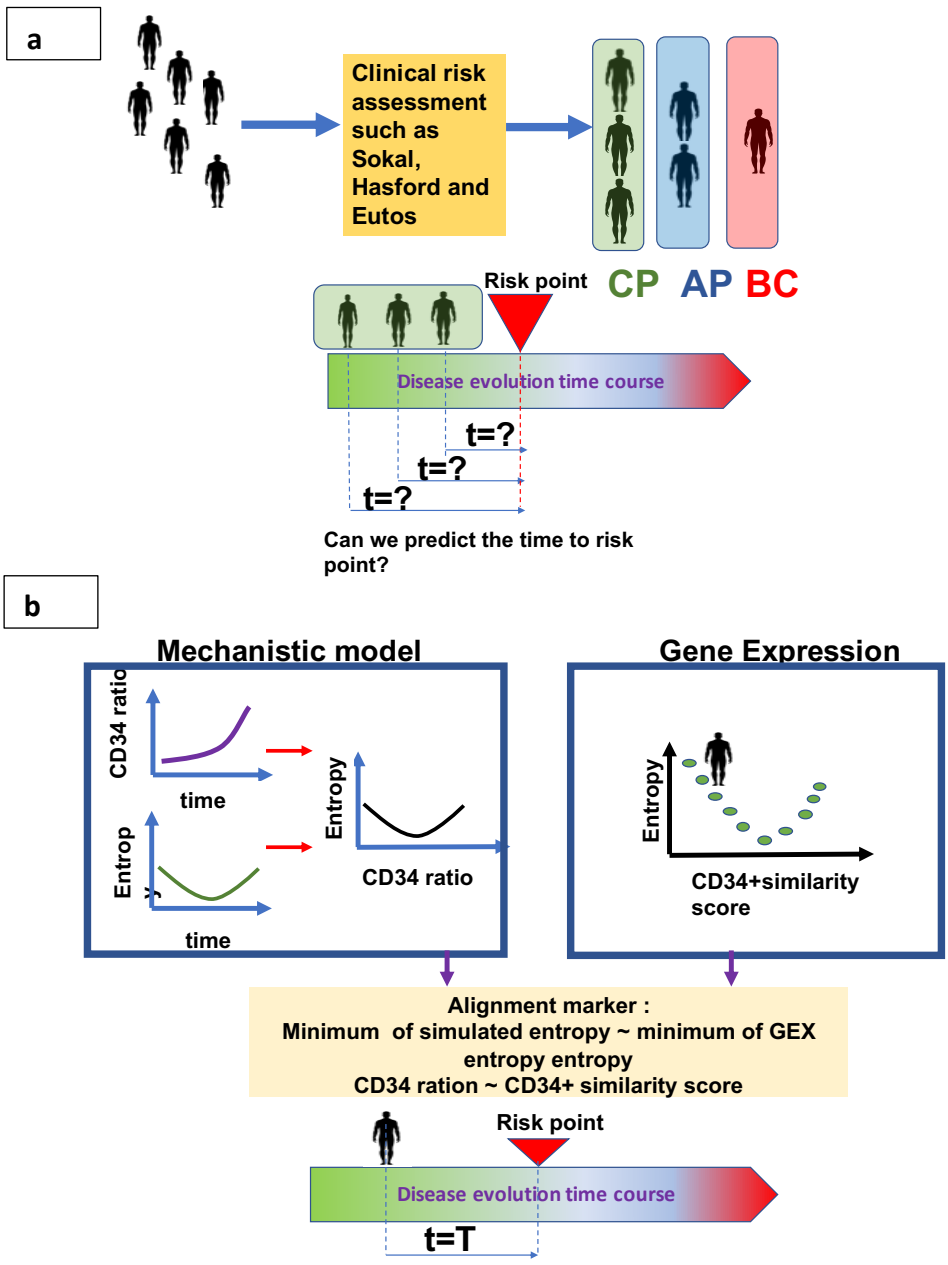


FIGURE 17 A: GOAL OF THE PROJECT. CAN A BIOMARKER PREDICT TIME TO RISK POINT IN CP-CML? B: THE WORKFLOW OF THE METHOD. INTEGRATION OF MECHANISTIC MODEL AND GENE EXPRESSION DATA FOR PREDICTION OF TIME TO RISK POINT

When projecting gene expression of the mixture of cells into PC space (figure 18b), we can see a clear trend of change from the chronic phase to blast crisis over an accelerated phase. However, when projecting gene expression of subpopulations into PC space (figure 18a), what comes more in the notice, is the separation of different subpopulations. Phase difference among each subpopulation is also observable, but the major and dominant effect is due to different cell types (figure 18a). This comparison confirms this assumption that changes we observe in gene expression of a mixture of cells, is in fact, a result of cell population shift. This result lets us integrate mechanistic population modeling with gene expression data for a better

individualized-risk assessment. Population dynamic models give us information about the time of disease evolution since the occurrence of the first mutation. Knowing this information about disease evolution time helps a lot for better therapeutic strategies as we can find out the time to the risk point. Knowing how to link the gene expression data and population dynamics model together, assessing gene expression alone can give us an estimation of time from disease initiation for each individual patient. To be able to achieve our goal, which is the integration of the gene expression and population dynamics model, we have to look for similar and correlated features, which assessable on both sides. In the following, we introduce these features and call them biomarkers from now on.

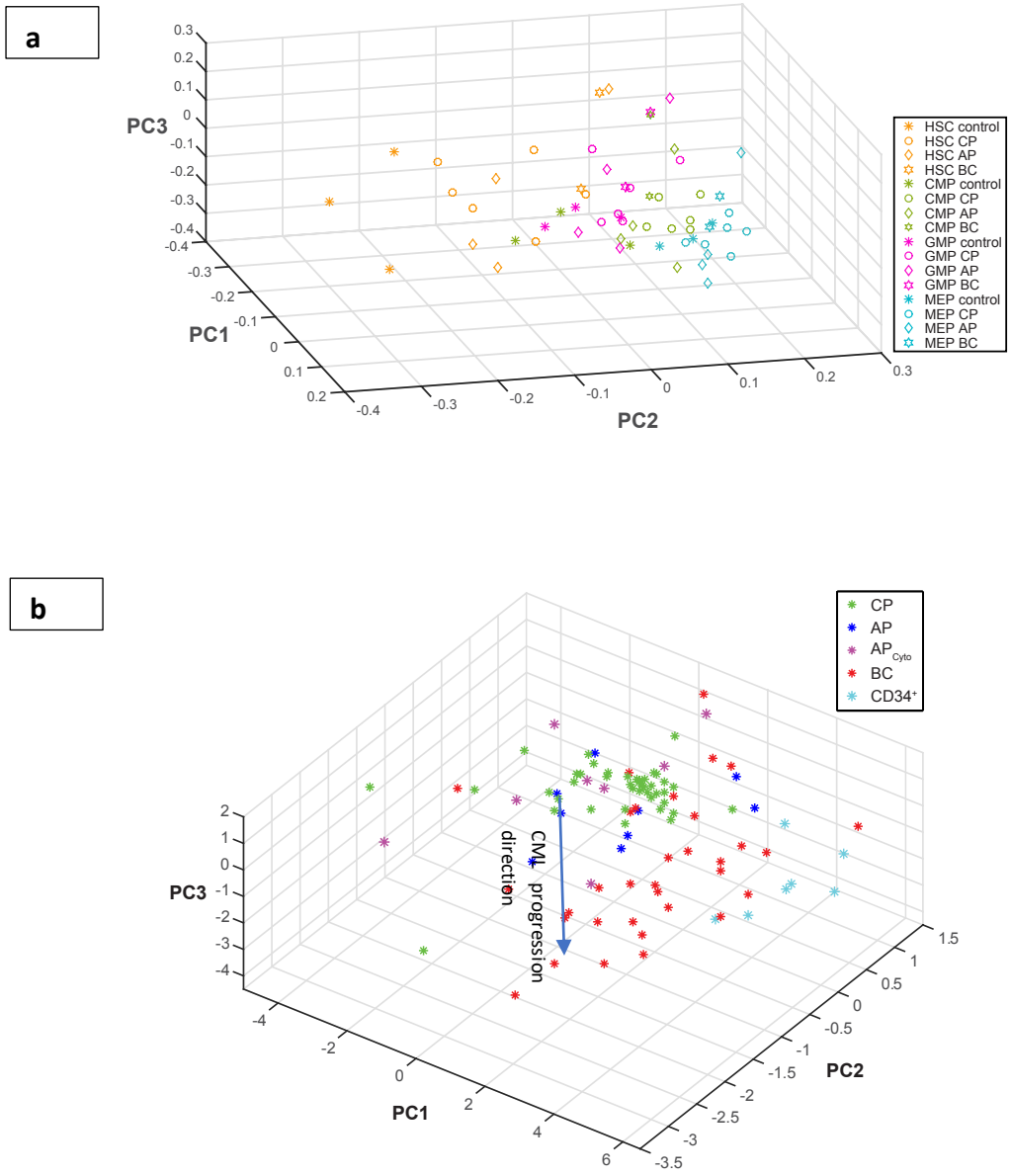


FIGURE 18: SHOWS THE DIFFERENCE BETWEEN GENE EXPRESSION IN DATASET GSE47927 (A) AND GSE4170 (B). TAKEN FROM [175]

4-2: PATIENT-DERIVED GENE-EXPRESSION BASED BIOMARKERS FOR CML DISEASE EVOLUTION:

4-2-1: CD34+ SIMILARITY SCORE

As mentioned previously, most of the invented clinical scores for CML classification and disease stratification cannot give information on disease evolution inside the chronic phase of CML. We look for a biomarker whose value plays a role as a risk alert to enter the accelerated phase while the patient is still in the chronic phase. Motivated by the previous studies,[176,177] which showed immature blast cells have a very similar profile as CD34+ and expansion of blast population is a sign of CML progression; we hypothesized that differentiation markers can indicate hematopoietic cell population changes associated with CML progression. In healthy hematopoiesis, the surface marker CD34 distinguishes immature cells such as HSCs and progenitor populations (CD34+) from more mature cells like committed precursors and differentiated cells (CD34−). Therefore, we developed a CD34+ similarity, score which captures how similar is the genomic profile of a patient to CD34+ cells.

4-2-1A: CALCULATING CD34+ SIMILARITY SCORE:

For calculating the CD34+ similarity score, we used the concept of PhysioSpace, developed by Lenz and colleagues[167]. The details of PhysioSpace concept are described more in details in chapter three (section 3-1-3-2-C). All of our gene expression-based biomarkers are assessed on GSE4170 dataset[168], which contains 42 samples in CP-CML, 9 samples in AP-CML, 8 samples in AP_{cyto} -CML, and 6 samples of normal CD34+ cells. We applied the PhysioSpace concept on the CML samples, considering CD34+ samples as the reference. We found out that the CD34+ similarity score not only distinguishes different CML stages but also adds additional resolution to CML progression inside the chronic phase for which existing conventional clinical parameters such as blast count fail (figure 19). Our CD34+ similarity score, therefore, is able to locate CP patients differently in the time course of the disease progression. This feature of CD34+ similarity score helps us further integrate (and *link*) gene expression profile of CML patients with cell population dynamics models that describe population change of healthy and cancerous cells in the chronic phase of the disease.

4-2-1B: CALCULATING CD34+ SIMILARITY SCORE BASED ON POPULATION DYNAMIC MODEL

To simulate the dynamics of cell population growth in both healthy normal hematopoiesis and CML hematopoiesis, there are plenty of models that we named and shortly described in the review part on CML modeling. We have chosen the model developed by Dingli and colleagues[125], because this model considers a compartmental hierarchy of hematopoiesis, consisting of 32 cell types, starting from stem cells in bone marrow, to the multipotent progenitor, precursor and fully differentiated cells in peripheral blood, while other models only consider stem cell and mature cells and no cell compartments in between. This hierarchy of hematopoiesis is described mathematically in a system of 32 ODEs. In general, they considered the same hierarchy and number of compartments in normal and CML hematopoiesis. What differentiates normal hematopoiesis form CML one, is the change in the differentiation rate, shown with ε. Based on their model, an active pool of 400 HSCs is responsible for normal marrow output with a differentiation property of $\varepsilon_0 \approx 0.85$. This differential property is equal across all compartments in their model. Self-renewal probability is, on the other hand,1 - ε. Both probabilities are considered constant across healthy hematopoiesis at a replication rate of $r \approx 1.26$ over approximately 31 divisions between the

HSC and the circulating compartments. For further details on the multi-compartment model of normal hematopoiesis, see the original publication by Dingli et al.[125] Based on their model, the number of active HSCs is constant. When a single mutation occurs in HSCs, the number of Healthy HSCs reduced by one, 400-1=399, and number of CML HSCs is increased by one, 0+1=1. However, in contrast to the constant amount of stem cells, CML patients do have increased counts of myeloid progenitors.

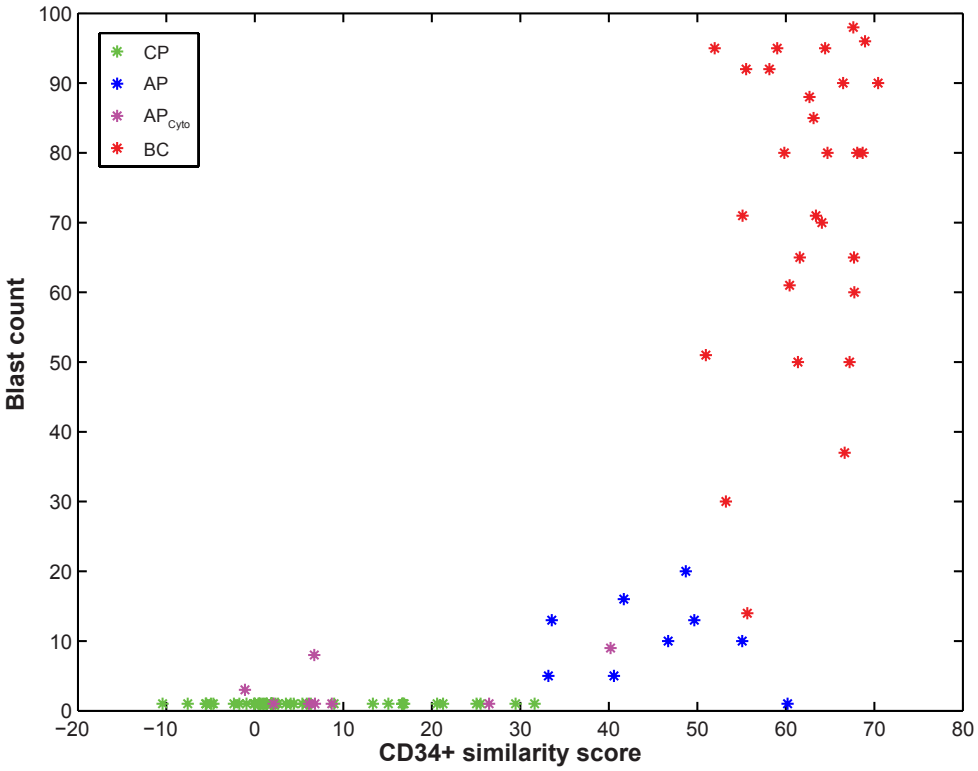


FIGURE 19: CD34+ SIMILARITY SCORE CAN DISTINGUISH CP-CML PATIENTS FORM EACH OTHER. TAKEN FROM[175]

CML is diagnosed when bone marrow output exceeds 10^{12} cells per day as a consequence of reduced differentiation probability, ε, in CML hematopoiesis. According to Dingli model, average clonal dynamics are approximated so that the number of cells in each compartment i ≥ 1 changes based on the following differential equation:

$$\frac{dN_i}{dt} = -d_i \times N_i + b_{i-1} \times N_{i-1} \qquad \textbf{Eq. (4)}$$

Where d_i and b_{i-1} are death and birth rate respectively, and both of them are functions of differentiation probability and replication rate:

$$d_i = (2\varepsilon - 1) \times r_i \text{ and } b_{i-1} = 2 \times \varepsilon \times r_{i-1} \qquad \textbf{Eq. (5)}$$

According to their model, $\varepsilon_{CML} < \varepsilon_0$, which means lower differentiation probability in CML cells. Starting with one single CML HSC, the disease lasts nearly 6 years until it becomes clinically evident at $\varepsilon_{CML} = 0.72$

HSCs and progenitor cells, due to their ability to differentiation, express CD34 surface marker (CD34+) in normal hematopoiesis. The precursor (from the 23rd compartment in [126]) and differentiated cells due to the lower ability of differentiation, lose CD34+ marker expression. Mature cells representing the majority of the blood are CD34−. During CML, however, a differentiation block causes population shifts to a pool of CD34+ leukemic blasts, shifting the ratio of CD34+/CD34− (CD34 ratio). This increase in CD34 ratio can be calculated according to the model by Dingli et al. as follows:

$$CD34_{Ratio} = (\sum_{i=1}^{31} N_{i,CML} + \sum_{i=1}^{23} N_{i,Normal})/(N_{32,CML} + \sum_{i=24}^{32} N_{i,Normal}) \qquad \textbf{Eq. (6)}$$

In order to integrate patient-driven CD34+ similarity scores and model-based CD34 ratio into a unique variable, we assume a linear correlation through a monotonically increasing function. This assumption is based on our observation that changes in gene expression are results of a change in cell populations (figure 18). As we do not know the exact location of each patient in the time course of disease evolution, we need another biomarker, which makes the mapping of the gene expression-based CD34 similarity score to model-based CD34 ratio possible. This new biomarker should have non-monotonic behavior along with the disease progression. Entropy's minimum/maximum can serve as an alignment marker. Entropy has been an interesting subject of tumor heterogeneity in many studies. Beretta and colleagues[178] showed that the cancer cell's transcriptome changes lead to measurable observed transitions of normalized Shannon entropy values (as measured by high-throughput technologies). In another study. Banerji and colleagues[179] proposed signaling entropy, as a biomarker to capture irregularities of genome-wide gene expression profile and to estimate intra-tumor heterogeneity.

In normal blood samples, as all the blood cells are healthy, the system is at its equilibrium, and the heterogeneity of cell populations is at its minimum. As a mutation occurs, cancer cells are born, and their clonal expansion disturbs the equilibrium of the system. As a result, entropy starts to increase. Before a certain point, these cancer clones are still in the minority in comparison to normal cells. As cancer clones expand, the system reaches a point at which the population of normal and cancer cells is equal. Here at this point, the heterogeneity and therefore entropy of the cell population is at its maximum. We call this point the phase transition point. After passing this point, cancer cells population dominate the normal cells. Therefore, the heterogeneity and as a result, entropy of the cell population is again decreasing. Thus, we assume that this maximum of cell population entropy can serve as a biomarker for phase transition[144]. In the following, we show the result of calculating entropy in both sides of our model: the simulated entropy from the mechanistic model and the calculated patient-derived entropy. For both calculation we used the formula of Shannon entropy:

$$\boldsymbol{S_k = -\int p(x_k) \log p(x_k)} \qquad \textbf{Eq. (7)}$$

4-2-2: SIMULATION OF ENTROPY FOR CELL POPULATIONS AND GENE EXPRESSION PROFILE BASED ON A MECHANISTIC MODEL

As described in the previous part, we assumed that at the phase transition point, the entropy of the cell population is at its maximum. Now, we want to not only simulate this entropy but also see how the cell population changes affect the entropy of the gene expression profile of the mixture of cells.

4-2-2A: SIMULATING ENTROPY OF CELL POPULATIONS

Based on the Dingli model, we can simulate the population change of 32 compartments of both CML and normal hematopoiesis. Overall, we have the population dynamics of 64 cell types and their respective ratio in a mixture of all 64 cell types. So, the density of different cell types at each time point can be considered as an arrangement of cell types in the blood sample of a specific CML patient. For each time point, and correspondingly a patient at this time point, we can calculate the entropy of cell population arrangement, based on equation 7, where k is the index for each time point or corresponding patient and $\boldsymbol{p}(\boldsymbol{x_k})$ is the density function of cell types in each time point (Figure 20 lower curve). It is worth mentioning that since Dingli Model is based on ODEs, the cell population dynamics and resulting calculated biomarkers, such as entropy of cell populations and entropy of gene expression, are continuous functions of time. However, since we apply numerical methods of MATLAB[180] to solve these equations, we have discrete time points. The representation of simulated entropies is, however, shown as a continuous function because of a considerable number of time points (precisely 1336) (figure 20).

4-2-2B: SIMULATING ENTROPY OF GENE EXPRESSION

We can also infer the entropy of gene expression of a mixture of cells based on the Dingli model. First, we have to simulate the gene expression profiles of CML disease evolution in the curse of time. If we consider a normal distribution of gene expressions for *a single cell type*, by multiplying the ratio of each cell type's population as the weight of how much a cell type can shift gene expression profile, we can simulate gene expression of the mixture of cells at each time point. As stated before, each time point can correspond to a specific patient located at that time point of the disease evolution. From the simulated gene expression profile, we can calculate the simulated entropy for each time point of the disease based on equation 7. With k is the index for each time point or each patient, and $\boldsymbol{p}(\boldsymbol{x_k})$ is the density function of the gene expression in each time point (figure 20 upper curve).

To put it in detail, as all of our calculations are done in MATLAB, to apply the Shannon entropy equation, we have to build a matrix as the gene expression dynamics over the CML evolution. Suppose that the implementation of the Dingli model resulted in the population value of each cell compartment in 1336 time points. We have overall all 64 cell types and their corresponding ratios. So, the result of the Dingli model is a matrix with 64 rows and 1336 columns. We call this matrix the Cell Ratio Matrix (CRS). We consider a random distribution of gene expressions for a cell compartment. To make it comparable with our patient-derived results in terms of the number of available genes, we consider ~6000 genes. Having 64 cell type ratios, simulated gene expression for 64 cell types in one single time point is a matrix of 64 rows and ~6000 columns. We call this matrix the Single Time Point Simulated Gene Expression Matrix (STSGEM). Having CRS and STSGEM, we can calculate the gene expression dynamics over the CML evolution. We call this third matrix simulated gene expression matrix SGEM. This is calculated as follows:

$$SGEM = STSGEM^T * CRS \qquad \textbf{Eq. (8)}$$

SGEM is a matrix with ~ 6000 rows and 1336 columns representing the dynamics of the expression of genes in a mixture of healthy and CML cells over the disease evolution time course. Now having SGEM, we apply equation 7 to calculate the Shannon entropy for each time point (corresponding to each patient). We can observe a V-Shaped curve in both of the

simulated entropies, however, here we can confer that as the entropy of cell population reaches its maximum (at maximum heterogeneity of cell types), which is the point that cancer cells take over the normal population, entropy of gene expression is acting conversely and reaches its minimum.

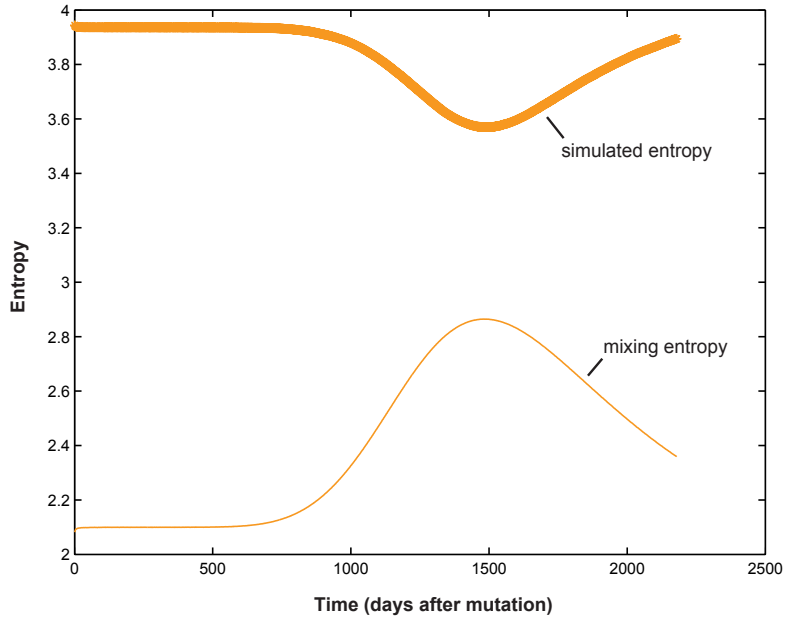


FIGURE 20: SIMULATED ENTROPY OF GENE EXPRESSION (UPPER CURVE), SIMULATED ENTROPY OF CELL POPULATION (LOWER CURVE). TAKEN FROM[175]

4-2-2C: PATIENT-DERIVED GENE EXPRESSION ENTROPY:

The data used in this section is GSE4170[168]. We consider that each patient represents a time point in the disease evolution. The matrix that we have to apply the Shannon entropy on has 91 rows (= 91samples and each sample corresponds to a time point) and ~6000 columns (genes). As we previously showed an increase of CD34+ similarity scores along with the disease progression, CD34+ similarity scores can quantify the location of each patient in disease evolution time. So, we can illustrate the dynamics of gene expression entropy over disease evolution (figure 21). Our objective here is to observe a V-shaped curve similar to our simulation observation. In order to confirm this V-shape curve, we did linear regression modeling using the 'fitlm' function with the quadratic term (MATLAB Linear Regression Package[181]). This resulted in a p-value of 0.014788 for the quadratic term, confirming the V-shaped behavior of the entropy in the patient's gene expression.

4-3: INTEGRATING GENE EXPRESSION DATA AND MECHANISTIC MODEL

Now that we assessed the corresponding biomarkers from both sides of our model, we are ready to lump them together. Our biomarkers are CD34+ similarity score and CD34 ratio as an indicator of the time in disease evolution spectrum and entropy, which its single minimum warns the phase transition point. To integrate our biomarkers, first, we have to make the CD34+ similarity score and CD34 ratio comparable and then to pin the minimum entropies

together. In order to match the CD34 ratio and the CD34+similarity score, we should first normalize them to put them in the same interval; therefore, we hypothesize that:

$$[\min(\text{CD34 ratio}), \max(\text{CD34 ratio})] \sim [t_0, 6 \text{ years}] \quad \textbf{Eq. (9)}$$

One way to compare the CD34 ratio and the CD34+similarity score is to normalize them in the same interval. Here we normalized both time surrogates to the interval of [0,1]. Based on our previous assumption of linear correlation between the model-derived CD34 ratio and patient-derived CD34+similarity score, we can observe similarities between the simulated and the patient-derived entropy. Now that both time surrogates are normalized and are within a comparable scale, we can quantitively map CP-CML samples to time of disease evolution provided by the model. To do the mapping, we have to find the corresponding normalized CD34 ratio value for each normalized CD34+ similarity. At first glance, that may seem easy and straightforward, but as mentioned before, because of the numerical solving of ODEs, we do not have a continuous spectrum of CD34 ratio, rather compact discrete values. Therefore, we calculated the distance between each patient's normalized CD34+ similarity score and all possible normalized CD34 ratios. Then we selected the nearest CD34+ ratio. Since the CD34+ ratio is a function of time, finding the corresponding CD34+ ratio is equivalent, finding the time point for that CP sample. Then, we *pinned* the minimum of simulated and patient-derived entropies together, such that each CP patient is at the minimum distance to its corresponding simulated data point in both independent scores (Figure 22b).

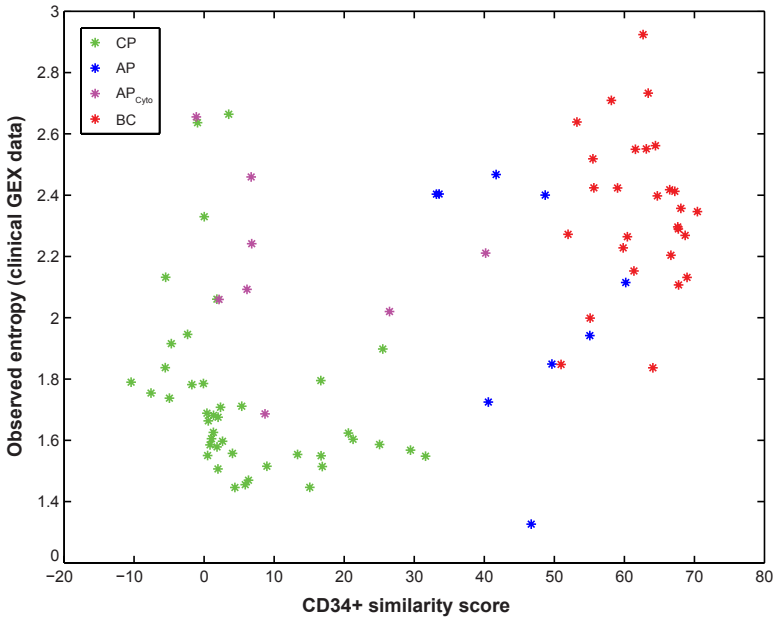


FIGURE 21: CALCULATING ENTROPY OF THE GENE EXPRESSION PROFILE OF CML PATIENTS IN GSE4170. TAKEN FROM[175]

4-4: PHASE TRANSITION IN CHRONIC PHASE

We hypothesized that the minimum of entropy serves as a biomarker for a phase transition. Based on this hypothesis, we divided samples into two phases of early CP (T1) and late CP (T2). After all alignment procedure, which is fully described in the previous part, the minimum entropy occurs at t=1476 days, almost 4.04 years after the occurrence of the first mutation in our simulations. This time point corresponds to CD34+similarty core of 0.403. It is worth mentioning that this number is critically dependent on the calculation of the CD34 ratio in

equation 6. Based on the multi-compartment model of the hematopoiesius[126], we considered the 23rd compartment as the threshold, after which in normal hematopoiesis, cells lose their CD34 marker expression. Testing the impact of parameter variations on this boundary, we calculated CD34+ ratios across all possible cut-offs for CD34+ expression and observed a robust T1–T2 boundary around 4 years. (figure 23)

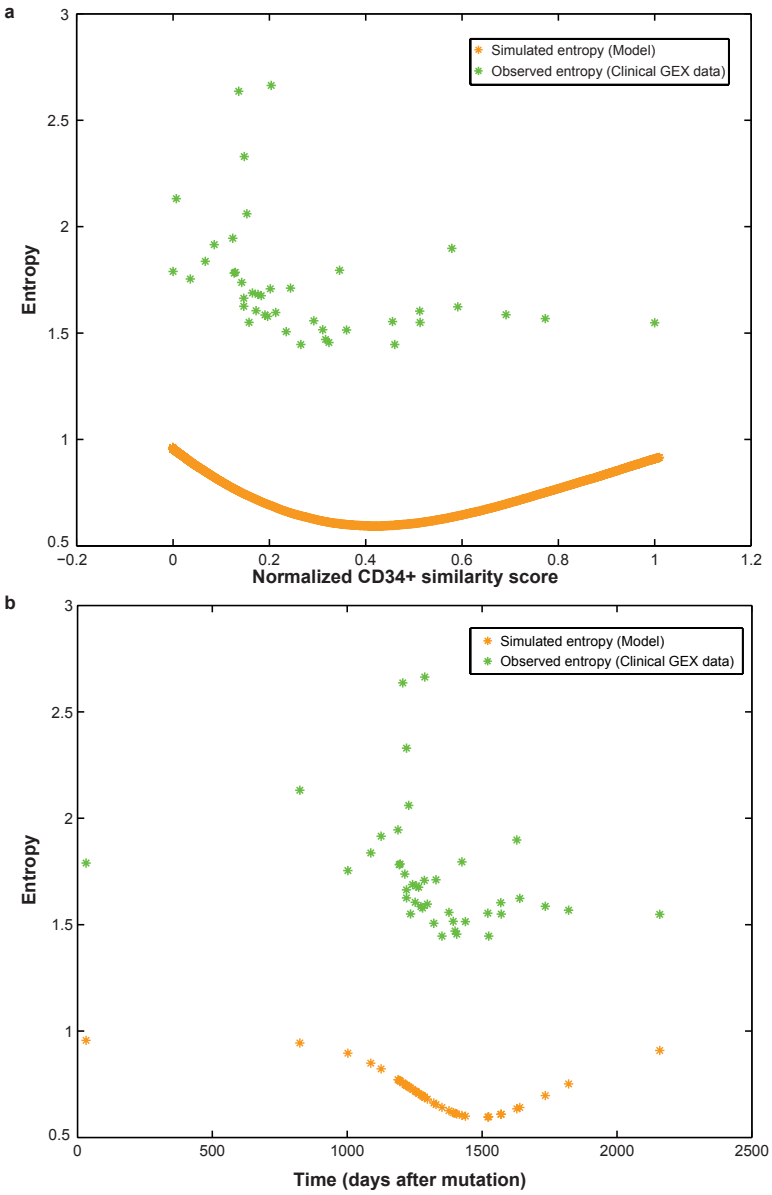


FIGURE 22: A) COMPARISON OF SIMULATED ENTROPY AND GEX-DERIVED ENTROPY. B) CORRESPONDING ENTROPY AND TIME FOR EACH PATIENT AFTER PINNING CD34+ SIMILARITY SCORE AND CD34 RATIO. TAKEN FROM[175]

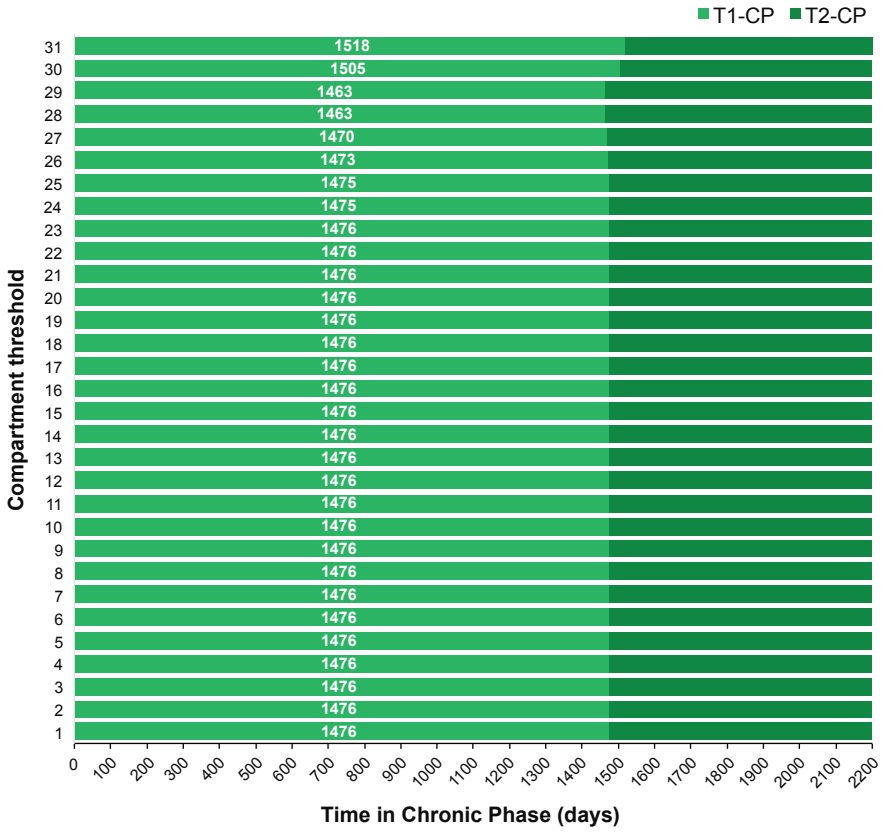


FIGURE 23 : EFFECT OF CHANGING COMPARTMENTAL THRESHOLD AS CUT OFF FOR CALCULATING CD34 RATIO. TAKEN FROM[175]

4-5: CORRELATION BETWEEN MODEL-DERIVED AND PATIENT-DERIVED BIOMARKERS

Now that we obtained corresponding CD34 ratio and CD34+ similarity score, we can assess our hypothesis on the linear correlation of these two scores. Applying linear regression model, we observed a significant correlation (R2 = 0.934, p = 3e-25) between corresponding CD34 ratio and CD34+similarity score (figure 24). We also checked the correlation of patient gene expression entropy with the simulated entropy (p = 0.0195). However, this correlation is not observable from figure 25a, and one may think it is specific to the existing samples. To confirm the independence of this correlation from our data set, we have done bootstrapping analysis, randomly dropping 10% of the data points in 1000 iterations without replacement. P-value frequency distribution for the corresponding correlations between the simulated and the observed entropies considering the remaining 90% of data points are shown in figure 25b. In 82.5% of the time, we achieved a p-value of less than 0.05 confirming the significant correlation between simulated and patient-derived entropies.

4-6: GENOMIC DIFFERENCES CONFIRM EARLY DISEASE PROGRESSION DURING CP CML

Now that we could identify two different stages inside chronic CML, we are interested to see if we can observe any genomic difference between these two stages, i.e., T1 and T2. For this purpose, we ran a statistical test between the thirty-four T1 samples and eight T2-CP samples

gene expression profiles. This resulted in the identification of 411 differentially expressed genes (FDR-adjusted p-value < 0.05). We also compared newly discovered stages with other CML phases, i.e., AP and BC to assess the significance of the difference between T1 and T2 in terms of their distance to further advanced stages. The distance here is defined as the number of differentially expressed genes. We can see from table 9 and figure 26, that as the disease progresses from T1 and T2, the number of differentially expressed genes between the disease state and AP/BC decreases. The number of genes differentially expressed between T1 and AP (66.8%), is almost double the number of genes differentially expressed between T2 and AP (32.3%). We can see the similar effect when comparing the number of differentially expressed genes in T1 vs. BC (80.7%) and the number of differentially expressed genes in T2 vs. BC (46.14%) These intense effects within CP suggest that T2-CP may represent a higher risk state with respect to progression. Distribution of log10 (p-value) of statistical tests between all possible combinations of stages is shown in figure 27. Again, since the number of samples inside T2-CP is very low, we did bootstrap analysis to confirm the robustness of these fractions (figure 28 and 29).

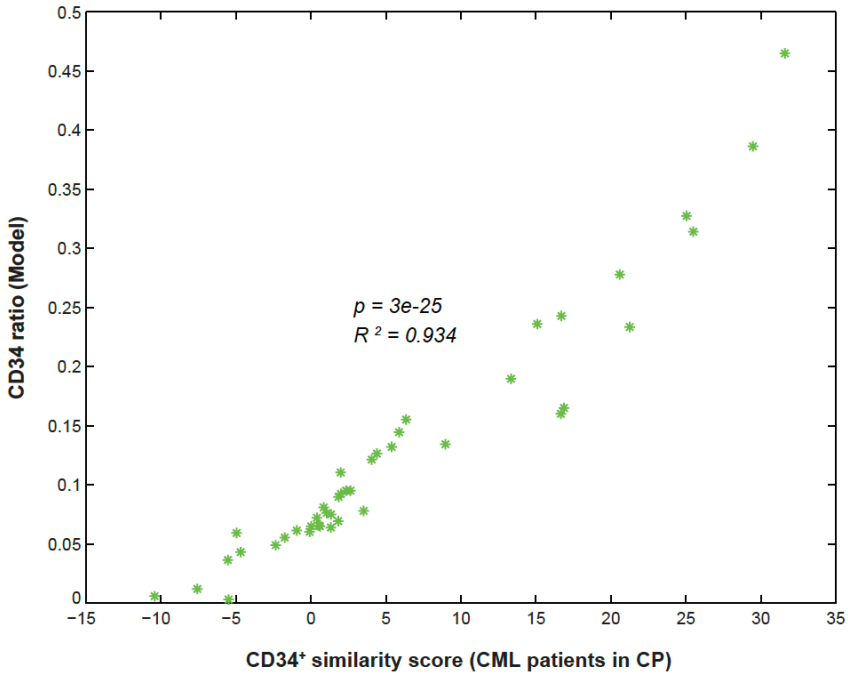


FIGURE 24: SIGNIFICANT LINEAR CORRELATION BETWEEN PATIENT DERIVED CD34+ SIMILARITY SCORE AND MODEL DERIVED CD34 RATIO. TAKEN FROM[175]

4-7: GENE SET ENRICHMENT ANALYSIS

In the next step, having 411 identified differentially expressed genes, we did a gene set enrichment ontology analysis to investigate biological ontologies with a significant difference between T1 and T2 stages. Using Panther online ontology analyzer[182], we found 163 significantly upregulated biological processes such as processes related to signal transduction involved in DNA damage, integrity checkpoints, or general cell cycle control. And 26 significantly down-regulated biological processes such as innate immune response and cellular differentiation and development. Several processes frequently observed among both up- and down-regulated such as protein translation, modification or trafficking, metabolism,

biosynthesis, or catabolic processes. The list of complete gene ontology analysis is brought as supplementary tables of our published paper of this study[175].

TABLE 9: PERCENTAGE OF DIFFERENTIALLY EXPRESSED GENES

Stages to be compared	%of differentially expressed genes
T1 vs T2	9.54%
T1 vs AP	66.8%
T2 vs AP	32.3%
T1 vs BC	80.7%
T2 Vs BC	46.14%
AP Vs BC	19.96%

TABLE 10: EXAMPLES OF SIGNIFICANT BIOLOGICAL PROCESSES

Upregulated Biological processes		Downregulated Biological processes	
Name	P-value	Name	P-value
Signal transduction in DNA damage	0.03	Innate immune response	0.02
Integrity check points	0.03	Cellular differentiation	0.006
General cell cycle control	0.04	Cellular development	0.007
...		...	

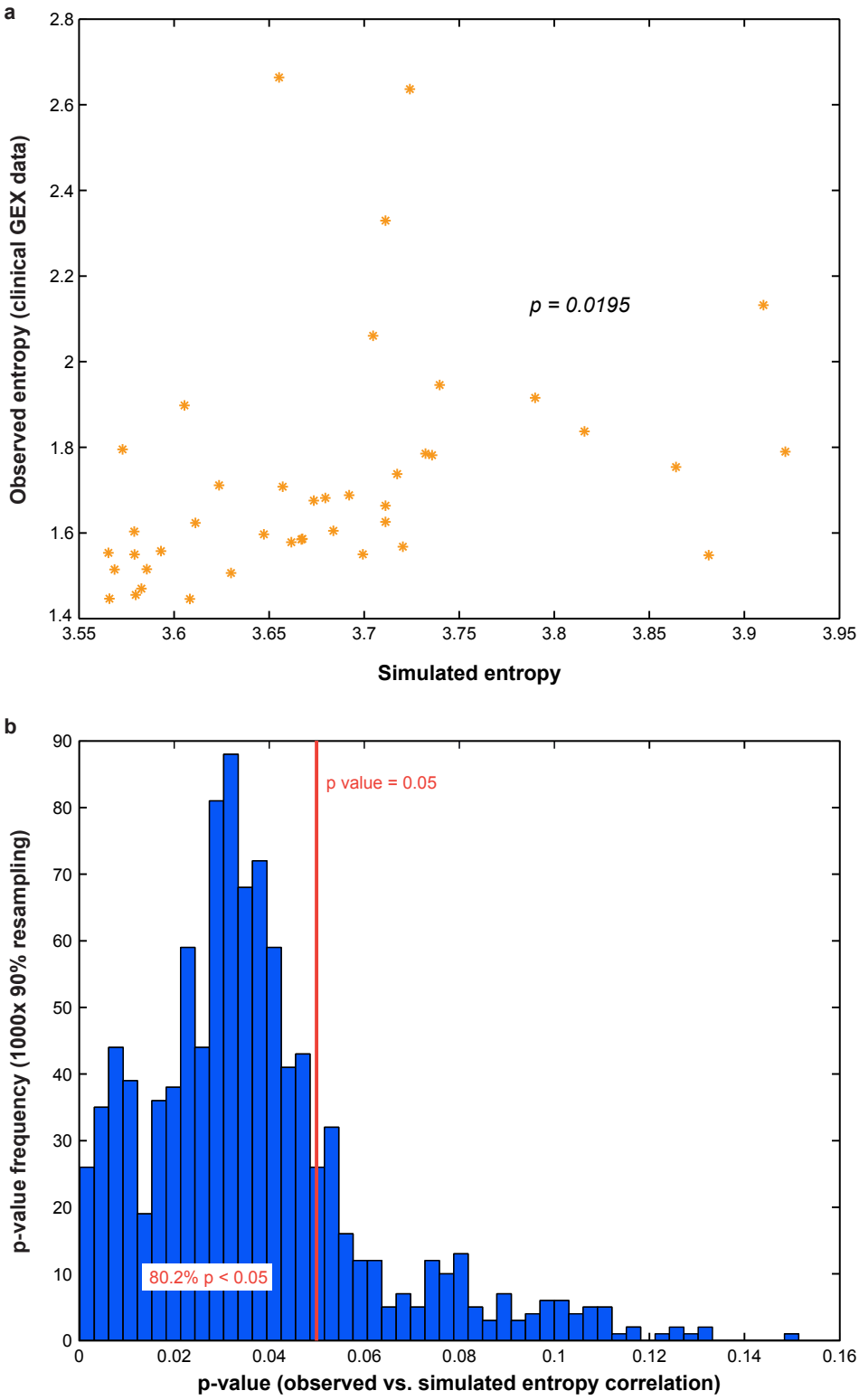


FIGURE 25 A) CORRELATION BETWEEN SIMULATED ENTROPY AND PATIENT- DERIVED ENTROPY, B) RESULT OF BOOTSTRAPPING ON P-VALUE OF LINEAR CORRELATION BETWEEN SIMULATED ENTROPY AND PATIENT DERIVED ENTROPY, TAKEN FROM[175]

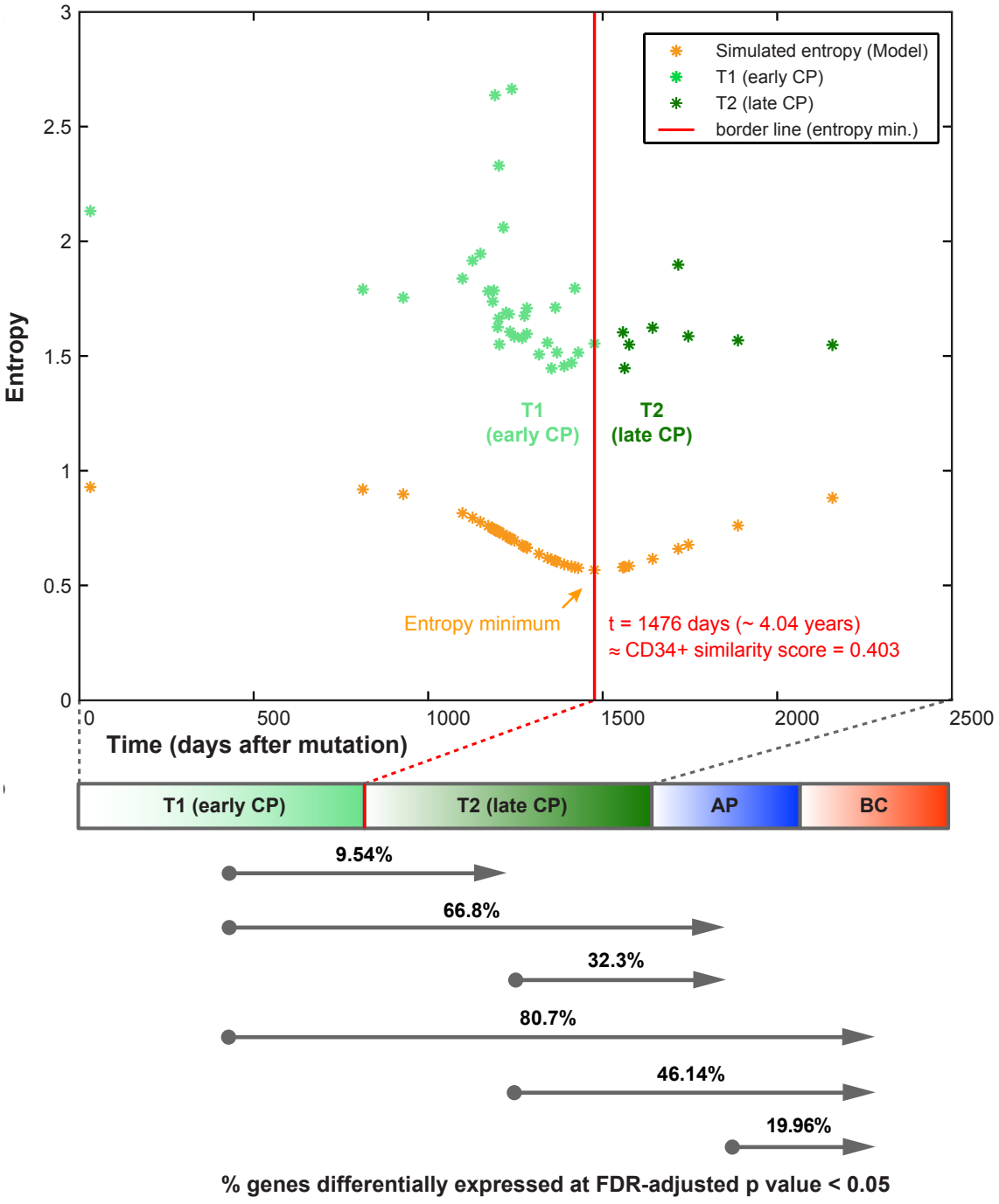


FIGURE 26: A) DIVIDING THE CP-CML PATIENTS IN TWO PHASES, T1 EARLY CP AND T2 LATE CP. B) THE PERCENTAGE OF DIFFERENTIALLY EXPRESSED GENES IN DIFFERENT COMPARISONS, TAKEN FROM[175]

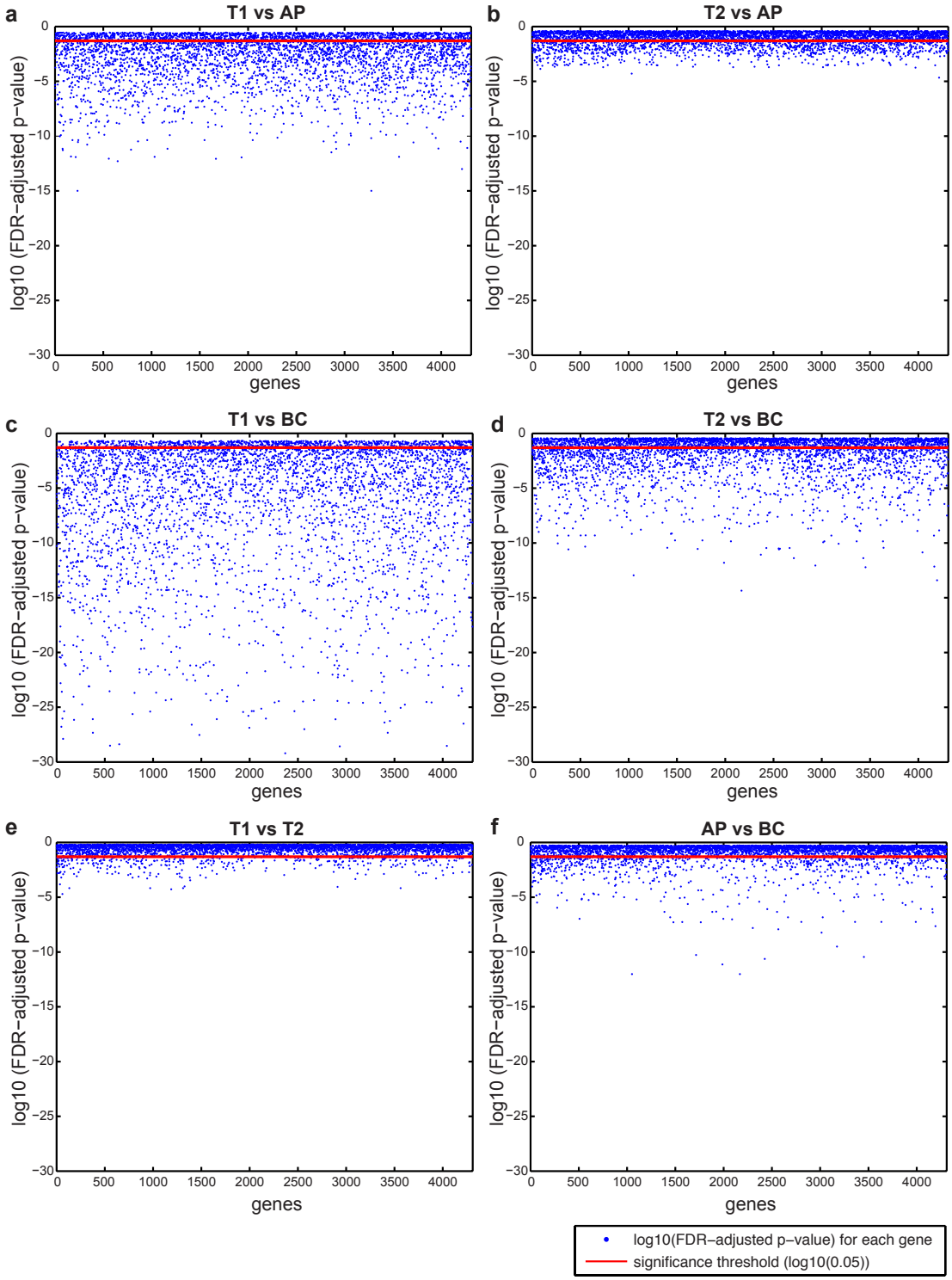


FIGURE 27SIGNIFICANCE OF DIFFERENTIAL EXPRESSION OF EACH GENE (BLUE DOTS) BETWEEN CML DISEASE STAGES. CUT OFF FOR SIGNIFICANCE VALUE IS 0.05 (FDR-ADJUSTED P-VALUE). T1: EARLY CP, T2= LATE CP, TAKEN FROM[175]

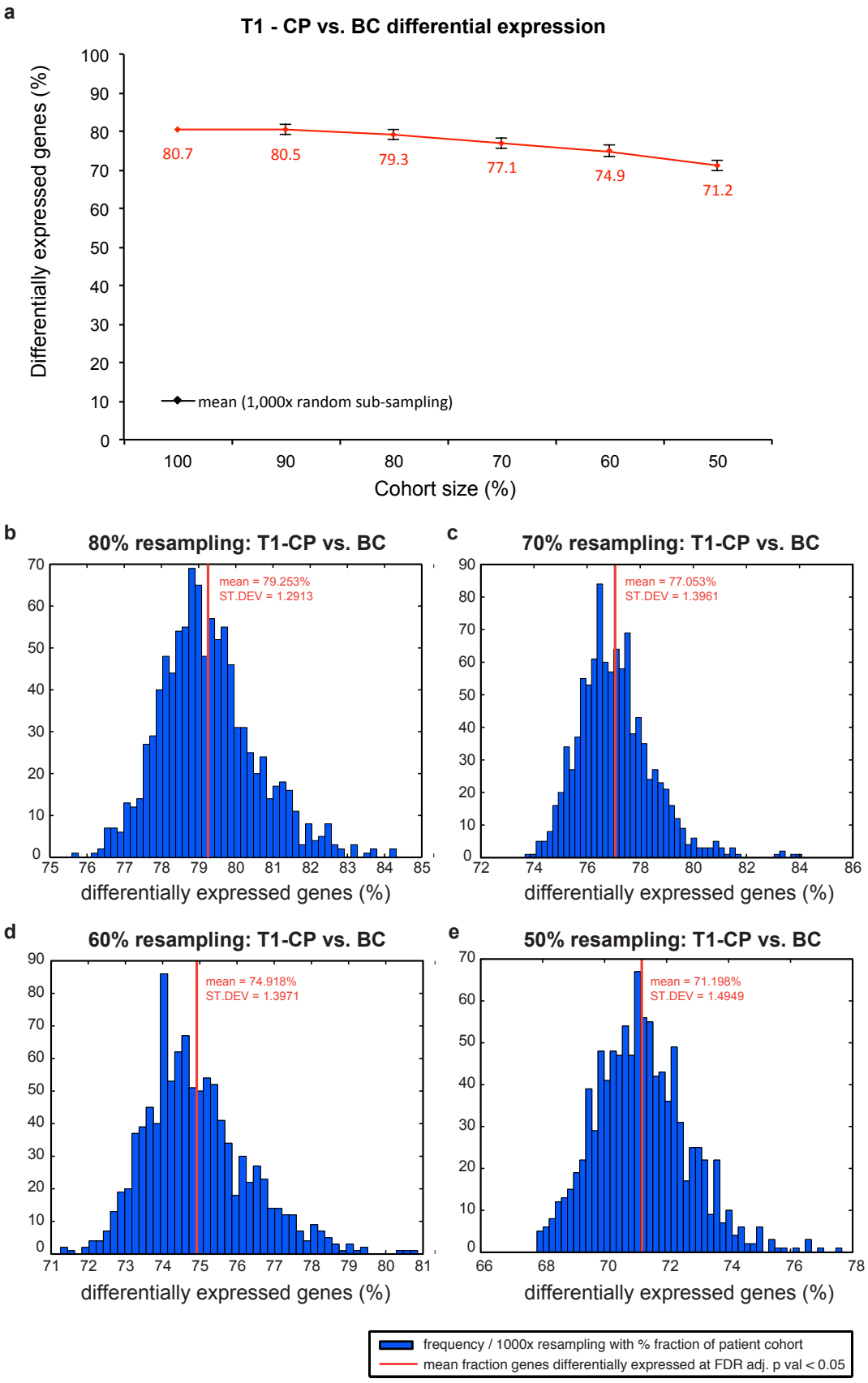


FIGURE 28: RESULT OF BOOTSTRAPPING WITH DIFFERENT RESAMPLING RATE ON PERCENTAGE OF DIFFERENTIALLY EXPRESSED GENES COMPARING T1 WITH BC, TAKEN FROM[175]

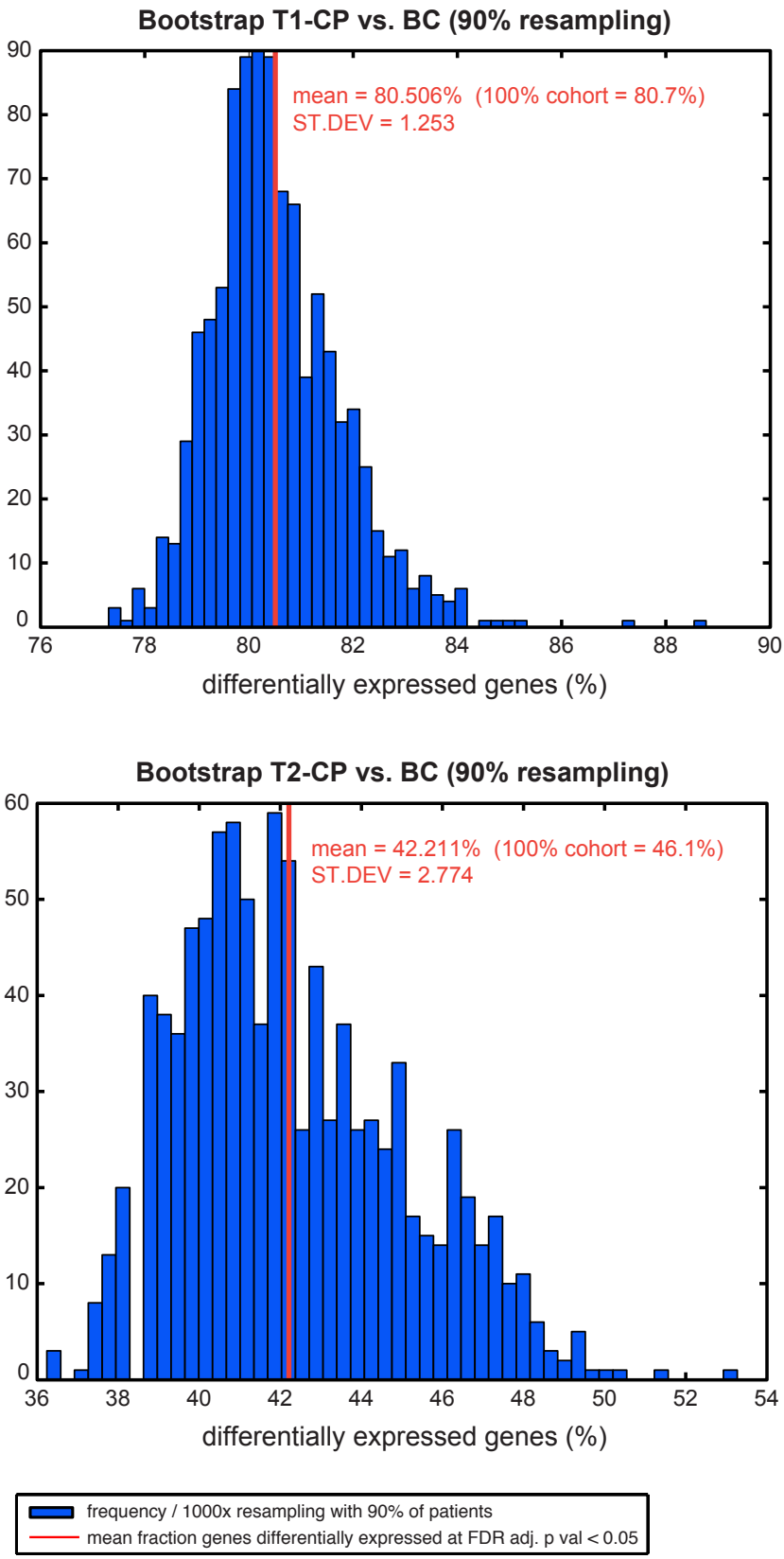


FIGURE 29: RESULT OF BOOTSTRAPPING AT 90% SAMPLING RATE ON PERCENTAGE OF DIFFERENTIALLY EXPRESSED GENES COMPARING T1 WITH BC AND T2 WITH BC.

4-8: CONCLUSION

In spite of significant development in the field of the molecular origin of CML and mechanistic models describing the progression of CML in terms of population dynamics, there is still a need for accurate risk assessment of patients in chronic CML, which is needed for better and more effective therapy strategy planning. There are still unanswered questions like how long before diagnosis, the disease has started, or how far is a CML patient from transition to accelerated phase? To answer these questions, we developed an approach by integrating heterogeneous sources of data, one from mechanistic model-driven simulations of blood cell populations and one from the gene expression profile of CML patients. Patients in CP share similar symptoms and phenotypes; however, not all of them are located in the same distance from phase transition point. We invented the CD34+ similarity score, which is able to distinguish CP patients from each other, what other conventional risk scores such as blast count fail to do. This patient-derived CD34+ similarity score is then shown to be linearly correlated with the model-derived CD34 ratio. Here we presented a novel approach that integrates the existing mechanistic models for CML progression with the gene expression data. The motivation for this integration comes from the idea that the observable changes in gene expression during disease evolution is due to the change in the arrangement of the cell populations. Motivated by the concept of phase transitions of complex systems in thermodynamics[183], we calculated the entropy of the simulated mixture of cell populations and the simulated gene expression profiles. We observed that where the cell mixture is at maximum heterogeneity, and therefore maximum entropy, the gene expression entropy is at its lowest value. We consider this minimum as an indicator for phase transition, which can be assessed both form patient gene expression data and mechanistic model. We used this feature of entropy as an alignment marker for mapping patient to disease evolution time course. Quantifying patient fractions within disease stages throughout both dimensions of entropy and CD34+ similarity score enables estimation of disease stage likelihood. As is depicted in figure 30, considering two groups of high and low entropy with four quarter divisions of CD34+ similarity score can give more resolution to patient stratification based on the gene expression profile of the patients. By placing a patient in a 2-D plane with entropy and CD34+ similarity score as building dimensionalities, we can introduce a new quantitative risk assessment method (figure 31).

We divided the entropy-CD34+ similarity score plane into eight main regions based on low-high entropy and quarter divisions of CD34+ similarity. Quantifying patient fractions within disease stages throughout both dimensions enables estimation of disease stage likelihood as: (figure 30)

- The 1st region with the first quarter of CD34+ similarity and low entropy:
  The probability of CP patients is very high in this region. 90% of patients in this region are in the chronic phase.
- The 2nd region with the first quarter of CD34+similarity and high entropy:
  With the same CD34+ score as the previous region, high entropy of gene expression profile can increase the probability of APcyto stage.
- The 3rd region with the second quarter of CD34+ similarity score and low entropy:
  While most of the patients again are in the chronic phase, but as we showed in our results, here, low entropy is an indication of phase transition.
- The 4th region with the second quarter of CD34+ similarity score and high entropy:
  We do not have data in this region to make any conclusion.
- 5th and 6th regions with the third quarter of CD34+ similarity score:

CP is found less frequently than expected randomly, and more patients than expected by chance are annotated as AP.

- The 7th region with 4th quarter CD34+ similarity score:
  No chronic phase patient is seen in this region and disease likelihood is divided between AP 30% and BC 70%.
- The 8th region with 4th quarter CD34+ similarity score:
  This region is characterized by 100% BC patients.

We have done bootstrapping analysis to support that this CML stage likelihood is not by chance (figure 32).

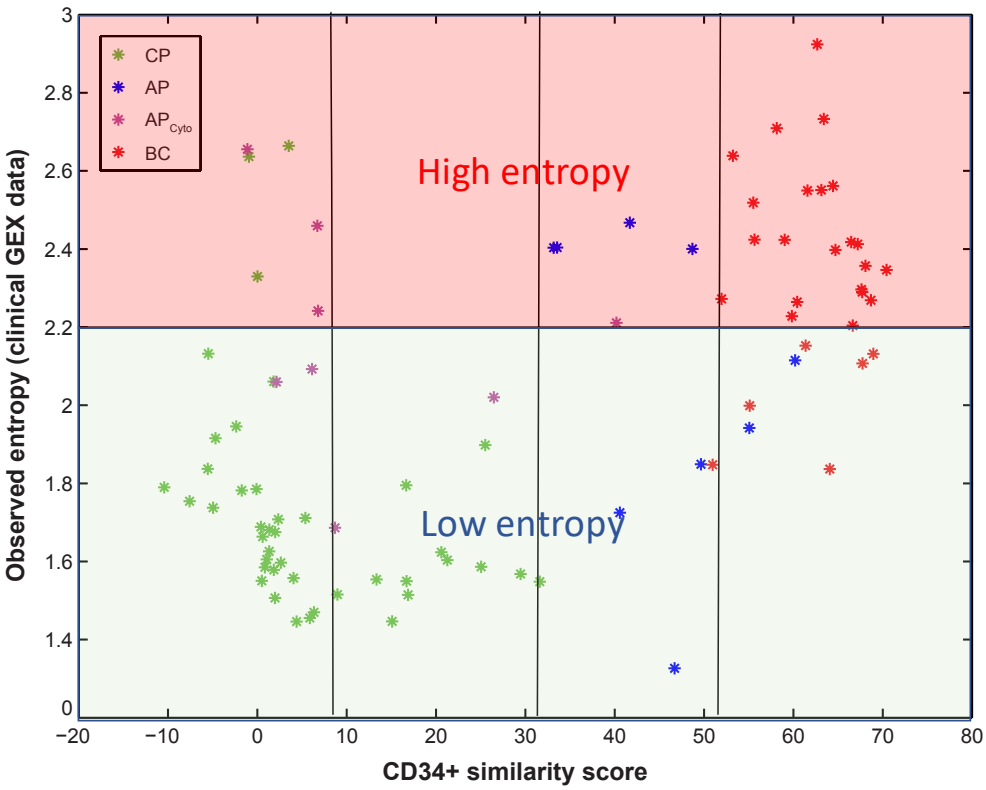


FIGURE 30: DIFFERENT REGIONS OF ENTROPY-CD34+ SIMILARITY SCORE PLANE

This quantification itself serves as a supporting rationale for genomics-based risk assessment in CML. In summary, our approach:

- relates patient status to disease time scale inferred from a population dynamics model,
- provides a new way to approximate progression risk, and
- reveals new insights into pathophysiological CML dynamics.

The combination of entropy and CD34+ similarity score enables a new sub-staging within the chronic phase of CML. We call these new subdivisions as T1, early CP, and T2 late CP. The statistical analysis resulted in 400 significantly differentially expressed genes out of ~6000 low-IR genes among CP subdivisions. Having this set of significantly differentially expressed genes, we did gene set enrichment analysis and found significantly different biological processes. Among singingly different biological processes, down-regulation of innate immune could be indicating impaired tumor surveillance. Other downregulated processes are cellular

differentiation and development, which can be a result of impaired hematopoietic differentiation implicated in proliferative advantages of transformed cells. We also found out processes, such as DNA damage, integrity checkpoint, and cell cycle processes to be upregulated, which is pointing toward responses to enhanced DNA damage and genomic instability, both implicated in disease progression and drug resistance[10]. Overall, reduced immune surveillance, impaired differentiation, and exacerbated genomic instability are hallmarks of cancer, reflected here by differential expression between two CP sub-groups.

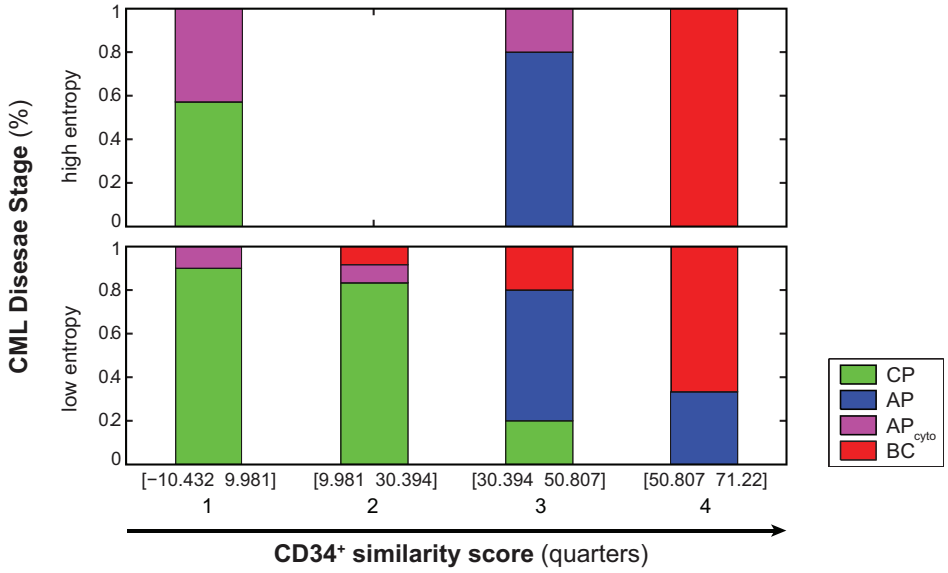


FIGURE 31: DISEASES STAGE LIKELIHOOD BASED ON ENTROPY-CD34+ SIMILARITY SCORE PLANE

Based on our calculation and mapping gene expression data on the time scale of the mechanistic model, the phase transition from T1 to T2 happens at a normalized CD34+ similarity of around 0.4, which equals to a time nearly 4 years after the occurrence of the first mutation. These numbers are limited to model parameters. Here we considered only one mutation to be the root cause of the expansion of blood cell populations. If we increase the number of mutations, the time interval of 4 years will decrease as the disease will develop faster [125]. In summary, our model enables quantitative interpretation of a combination of conventional and genomic markers in terms of disease evolutionary time. This integration is capable of being extended in other clinical research, where no mathematical or predictive biomarkers exist to date.

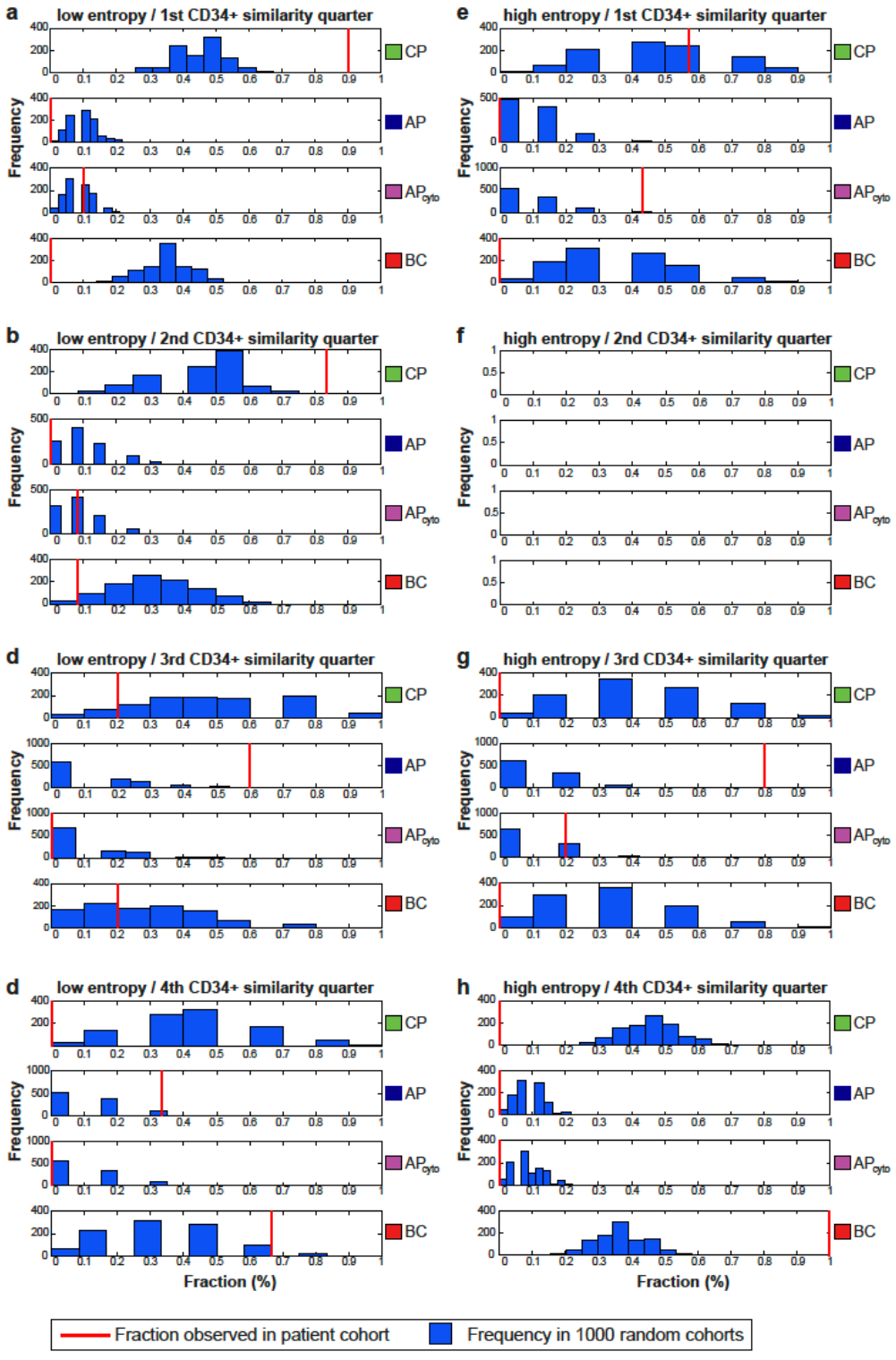


FIGURE 32: BOOTSTRAPPING RESULT FOR EACH DISEASE STAGE LIKELIHOOD (8 PANELS A-H). IN EACH PANEL Y AXIS IS FREQUENCY AND X AXIS IS THE PROBABILITY OR FRACTION OF EACH STATE (EACH RECTANGLES).

CHAPTER 5: DEVELOPING MPN CLASSIFIERS WITH GENE EXPRESSION AND CLINICAL DATA

In this chapter, we focus on investigating disease progression patterns in classic MPNs i.e., ET, PV, and PMF. To overcome the complexity of the diseases, we aim to model the disease progression at different levels of data: laboratory clinical data and genomic data.

As we already addressed in chapter 3 and in the materials section, for analysis of Philadelphia negative MPNs, we have two types of data from heterogeneous sources: gene expression data GSE26049 and clinical data GSG-MPN. As our modeling strategy is based on the source of the data, we have developed two different workflows.

First and for the gene expression data, we are interested in finding differences in genomic levels. The workflow is as follows:

- Applying PCA:
  As we are dealing with high dimensionality (number of genes) in our data, we start with dimension reduction by applying PCA.

- MPN Classification based on the PCs of gene expression:
  Applying different machine learning algorithms for finding out the best subsets of PCs, which can classify different MPNs.
- Developing a biomarker-based on the selected PCs:
  Based on the results of the previous step, we used linear regression to develop a biomarker to capture MPN progression.
- Improvement of the classifier by adding information from PhysioSpace:
  In this step, we investigated if adding physiologically related information to each sample can improve the quality of MPN classifications.

Secondly, for the GSG-MPN dataset, we try to develop a hybrid model by developing black box sub-models on separate partitions of the whole feature space, which are cell count and non-cell count data. The workflow of the hybrid model for GSG-MPN data is as follows:

- Development of the first sub-model:
  Training a classifier for MPN stages based on the cell count data.
- Calculate the misclassification score of the cell count classifier.
- Development of the second sub-model:
  Training a regression model on the non-cell count data for prediction of the classifier misclassification error.
- Integration of the outputs of the two sub-models:
  Training a classifier by using the predicted misclassification error from the second sub-model and predicted classes from the first sub-model for the improvement of classifications.

The motivation of our research in this part is, in fact, that as the assessment of genetic data is not a common procedure in clinics, there is a need for a precise biomarker based on clinical covariates that are capable of capturing the genotypic characteristics.

5-1: GENE EXPRESSION ANALYSIS

Due to the presence of JAK2V617F mutation in all three types of MPNs, there is a hypothesis suggesting that the phenotype of patients with JAK2V617F-positive ET resembles the early stages of PV[52]. From the other side, since that ET and PV sometimes develop to myelofibrosis, there is s hypothesis that ET, PV, and MF can be considered as different phases of the same disease[52]. Considering this hypothesis, we are looking for a biomarker that is able to track the progression of this disease from ET over PV to PMF. To analyze the gene expression data in this view, we firstly apply dimension reduction strategies such as PCA and PhysioSpace to track changes and see if there is, in fact, a continuum of change from ET to PV.

5-1-1: DIMENSIONALITY REDUCTION OF GENE EXPRESSION ARRAYS

As we deal with high-throughput arrays of gene expression, the high dimensionality of the data raises some challenges for observing disease progression trend. As genes are interacting with each other in a network, many of them are correlated with each other. To reduce the dimensionality, we apply PCA. In this section, we used the GSE26049[184] dataset. GSE26049 contains expression profiling by array of human whole peripheral blood. This makes it perfect for our study since we are looking for a mixture of cells and that is based on our results in CML that the changes in gene expression are, in fact, the results of the changes in population ratio of different cell types (section 4-1, figure 18). GSE26049 contains 21 healthy control samples, 19 ET patient samples, 41 PV patient samples, and 9 PMF patient samples. The number of genes in the dataset is 17275. We applied PCA on the full 17225 genes (details of complete procedures are brought in the appendix, section A1).

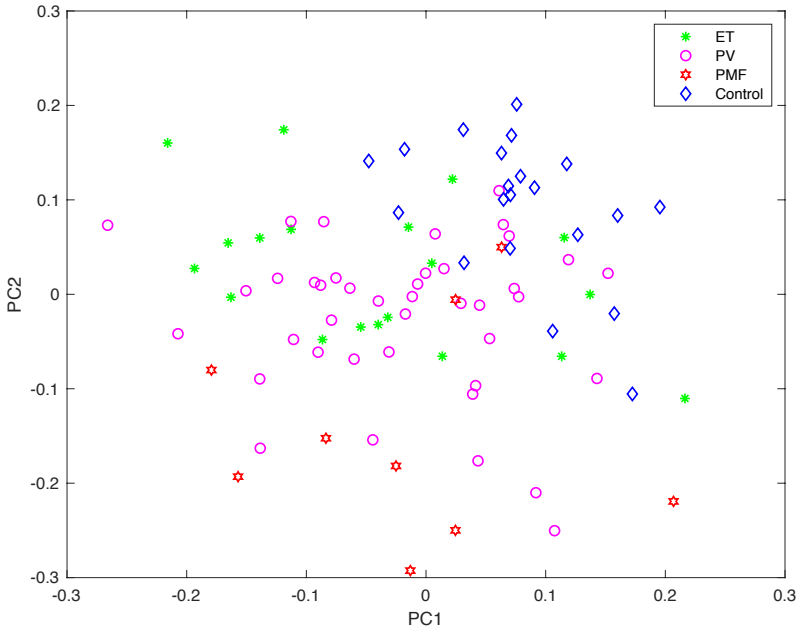


FIGURE 33: FIRST AND SECOND PCS OF GSE26049 ARE SHOWN HERE AS A RESULT OF PCA.

As indicated in figure 33, which shows PC1 vs. PC2, there are no obvious clusters of different MPN diseases at the gene expression level. And in all covered region of PC space, all most all three diseases are overlapping such that even there are control samples which are located among ET and PV regions. These overlapping phenomena support previous studies that MPN

disease share similar characteristics[185]. However, we can see a direction of change from control to PMF via ET/PV. Most of the control samples are located upright of the PC1-PC2 region, and most of the PMF samples are located down left of this space, while most of the ET and PV samples are located in the middle region. This indicates that there might be a continuum of disease progression from the healthy state to PMF.

5-1-2: DEVELOPING A CLASSIFIER BASED ON PCS

PC1-PC2 plot in figure 33 would suggest that there might be a progression trend from control to PMF. To capture this trend, we tried different techniques, such as logistic regression, SVM, and NN. To choose the right technique and the best subset of PCs, we run the cross-validation on with 70% of data samples as training and 30% as the test set for 100 repetitions. In the appendix (section A1-1: A1-3), the procedure of how to choose the best set of PCs and best modeling algorithm is shown and fully described in detail. Based on the cross-validation results (in the appendix, section A1-4), logistic regression performs as the optimal MPN classifier with the first 4 PCs as its input. The final accuracy is about 83.3%. In detail, the accuracy for each class of MPN based on a logistic regression classifier build upon the first 4 PCs of gene expression is reported in table 11.

TABLE 11: ACCURACY OF PREDICTED CLASSES FROM LOGISTIC REGRESSION CLASSIFIER ON GENE EXPRESSION ARRAY

MPN Stage	Accuracy of classification
ET	47.4%
PV	92.7%
MF	77.78%
Control	100%

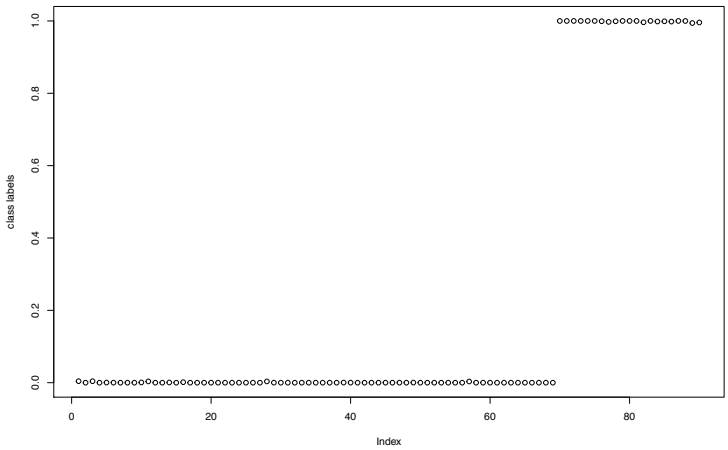


FIGURE 34: ALL CONTROL SAMPLES, ARE CLASSIFIED CORRECTLY. THE X-AXIS IS THE PATIENT INDEX, FROM 70TH PATIENT TO 90TH ARE CONTROL SAMPLES. AND NONE OF THE NON-CONTROL SAMPLES (INDICES FROM 1TO 69) CLASSIFIED AS CONTROL.

5-1-3: SETTING UP A PROGRESSION BIOMARKER WITH LINEAR REGRESSION

Now, that we have seen the first four gene expression PCs can perfectly classify control vs. non-controls and based on our observations from figure 33, that there might be a progression from control to PMF via ET/PV, we can set up a biomarker with linear regression. This one-dimensional biomarker can help us for further analysis based on PhysioSpace. In the linear regression model, we have to solve the following equation:

$$X * R = Class$$

Or

$$class_i = \alpha + \beta PC1_i + \gamma PC2_i + \theta PC3_i + \rho PC4_i \qquad \textbf{Eq. (10)}$$

Where *X* is a matrix containing the first four PCs, *R* is the regression vector, and *Class* is a vector of samples classes, here we have two classes: controls and non-controls. *i* is the index of each patient and *R*= $[\alpha, \beta, \gamma, \theta, \rho]$. The unknown in the equation above is the *R* vector. By solving the equation above we obtain:

$$R = [0.77, -1.73, -1.85, 0.62, -1.36]$$

Now by having *R*, we can calculate new scores for each sample by multiplying R and PC matrix, X. The box plot of new scores is shown in figure 35.

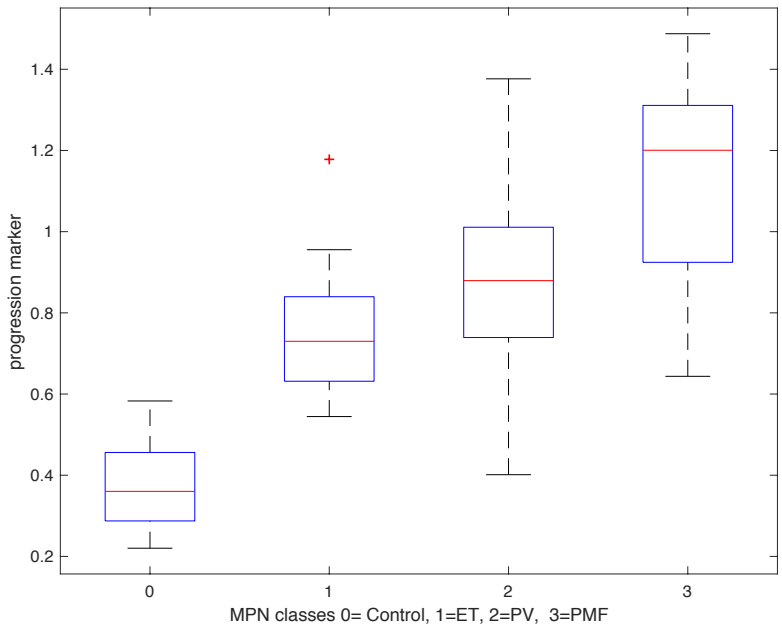


FIGURE 35: BOX PLOT OF PC-BASED PROGRESSION MARKER IN DIFFERENT STAGES OF MPNS. IN EACH BOX, THE CENTRAL MARK INDICATES THE MEDIAN, AND THE BOTTOM AND TOP EDGES OF THE BOX INDICATE THE 25TH AND 75TH PERCENTILES, RESPECTIVELY. THE WHISKERS EXTEND TO THE MOST EXTREME DATA POINTS NOT CONSIDERED OUTLIERS, AND THE OUTLIERS ARE PLOTTED INDIVIDUALLY USING THE '+' SYMBOL.

As we see in figure 35, values of this biomarker are overlapping between different MPN stages; however, an increment of the medians is observable as the stages change from control to PMF via ET/PV. If we consider a continuum of progression from ET/PV to PMF, then this progression marker can serve as the time of disease progression, when the whole systems of

genes are moving from healthy states toward PMF as the most progressive state. The statistical test resulted in a significantly different distributions of the progression marker among different MPNs (table 12). As MPNs are blood diseases, with abnormal blood counts, and based on our previous studies that change in gene expression profile is, in fact, the result of the change in population shift[175], we are interested in changes of gene expression in terms of different cellular entities. To investigate how the population of different cell types change the gene expression of the mixture of cells, we apply PhysioSpace.

TABLE 12: P-VALUE OF T-TEST ON DISTRIBUTION OF PC-BASED PROGRESSION MARKER

Comparison	P-value of t-test
ET vs. PV	0.0117
ET vs. PMF	0.0000
ET vs. Control	0.0000
PV vs. PMF	0.0000
PV vs. Control	0.0000
PMF vs. Control	0.0000

5-1-4: INVESTIGATING OF GENE EXPRESSION BY PHYSIOSPACE

In this section, we aim to discover the behavioral change of MPNs in the gene expression level with the help of physiology-related dimensions of PhysioSpace[167]. The details of calculating scores across new dimensions (PhysioScores) are fully described in chapter 3.

5-1-4A: PROJECTING MPN DATASETS ON PHYSIOSPACE.

We have in general three datasets. GSE26049[184], which is a public dataset available in GEO omnibus and contains MPN samples as described before. The second dataset, GSE120362[186], is from a study done by Czech and colleagues from Uniklinik Aachen. For comfort sake, I call this dataset from now on as the IZKF dataset. The third dataset is actually the PhysioSpace created by Lenz and colleagues. It is a 2-dimensional matrix with 369 physiology axes and 12123 genes.

TABLE 13: DATASETS USED FOR PHYSIOSPACE ANALYSIS

Dataset	Number of samples
GSE26049	90 MPN and healthy samples: 19 ET, 41 PV, 9 PMF and 21 healthy samples
GSE120362	32 MPN samples 6 ET, 10 PV, 9 PMF samples and 3 Secondary myelofibrosis samples, 4 healthy samples
PhysioSpace	Build upon more than 5000 samples in 369 different physiological groups

To make the mapping and comparison between these heterogeneous sources of data possible, we should extract common genes. However, before that, since all the datasets have different gene annotations (Affymetrix ID, Gene symbols, and Entrez ID), using "org.HS.eg.db"[187] packages in R, we unified all the gene identifiers to Entrez ID. Now having the same identifiers, we extracted 11597 common genes between all three datasets. One unique feature of PhysioSpace is that the distance to the reference is not dependent on the array types or normalization protocols of each dataset[167]. Therefore, we just use the Z-score normalization on each dataset separately. Another crucial step in calculating PhysioScores is the definition of a reference. For both the GSE26049 dataset and the IZKF dataset, the reference is calculated as the arithmetic mean of healthy control samples in each dataset. Also, for a better interpretation of the results, it is important to define a meaningful reference for PhysioSpace. The original reference in PhysioSpace is an average overall of ~5000 samples. These samples also include non-related blood tissues or samples, such as samples from the brain, heart, skins, etc. This makes this virtual reference hard to interpret in terms of disease progression in MPNs, so it makes sense if we shift the reference to a blood-related tissue such as whole healthy blood tissue or HSC.

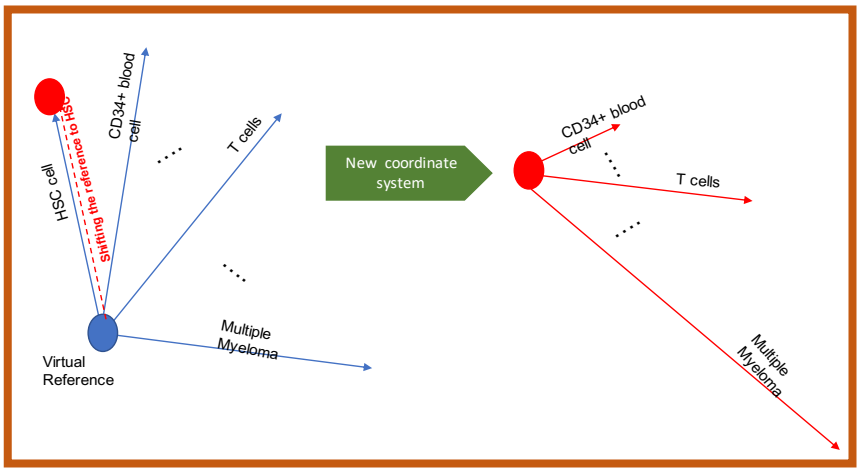


FIGURE 36: SHIFTING THE ORIGINAL VIRTUAL REFERENCE OF PHYSIOSPACE TO BLOOD-RELATED REFERENCE.

Calculating the most upregulated and downregulated genes is the next step. It is worth to mention that here to calculate the distance from reference, we calculated the fold change of expression for each gene, and because expression values in GSE26049 and IZKF datasets are in logscale, fold change is a subtraction of gene expression in different states. For each gene, the higher the absolute value of the difference, the more significantly differentially expressed is that a gene between two states.

The final step of calculating PhysioScores is the comparison of rankings of the two sets of 579 genes with rankings of genes in each of 369 PhysioSpace axes. This means to find the similarity between our datasets and each physiological axis. Here in our research, since we are interested in MPN as blood disorder, we calculated PhysioScores only for blood-related tissues or cell lines such as blood, HSCs, monocytes, macrophage, etc. The complete list of investigated PhysioScores is brought in table 14.

As we have many Physio axes, PhysioScores in the axes of the blood cell types and some blood-related disorders are brought as supplementary figures (Sfig1-Sfig18). Some of the results that are more interesting in the sense of disease progression from control to PMF via

ET/PV are shown figures 37 and 38. Since we have previously[175] shown that gene expression profile is a mirror of population shift, firstly, we focused on comparing PhysioScores of blood cell types among different MPN stages. Our choices are restricted to the existing axes of PhysioSpace. Existing blood cell axes are T cells, B cells, lymphocytes, leukocytes, monocytes, macrophages, HSCs, CD34+ blood cells, plasma cells. There is also an axis considering blood as a tissue itself. There are also some other interesting axes such as primary fibroblast, which can be indicative of differences between PV/ET and PMF. Based on the previous studies[168] and our research in CML modeling, as a result of the differentiation block during the CML evolution, more developed phases of CML show more stemness by expressing CD34+ markers. So, it is also interesting to study the stemness behavior of different MPNs. For this purpose, we calculated PhysioScores on axes such as adult progenitor cells and embryonic stem cells. It is also interesting to compare similarities/differences between MPNs and other blood cancers such as ALL and multiple myeloma.

5-1-4B: RESULTS OF PROJECTION OF GSE26049 INTO PHYSIOSPACE

Depending on the stages of MPN disease, we observe different phenotypes along different dimensions of PhysioSpace. In all the figures in this section, PhysioScores are depicted against the PC-based progression marker. Before investigating in detail, the behavior of each stage in different PhysioSpace dimensions, the effect of shifting the reference of the coordinate system to either whole blood or HSC is shown in figure 37. As it is obvious in figure 37, by reversing the direction, the signs of PhysioScores are also reversed. In figure 37a, a positive score means, as we are moving away from the blood, we are nearing to HSC tissue. PMF has positive scores, which means blood biopsy samples from a PMF patient show higher similarity to HSC in comparison to a healthy blood tissue. While in figure 37 b, we see that along the positive direction, which is moving away from HSC toward blood, PMF samples show negative scores, meaning that they are moving in a reverse direction (they go from blood to HSC) and they show similarity to HSCs.

In the following, the results of the projection of the MPN sample to Physio space axes are investigated in different disease categories.

Results in ET:

ET has the most interesting behavior among all other MPN types, and that is because in most of the PhysioSpace axes, there is a monotonic behavior observable among ET samples across the PC-based progression marker (supplementary figures and correlation coefficient >0.5 in table 15). Depending on the reference of PhysioSpace, however, results are different. When blood is considered as the reference, 47 axes out of 66 scores such as B cells, macrophages, T cells, lymphocytes, erythrocytes, HSCs, CD34+ blood and etc., have high negative correlation (<-0.5) with progression marker (table 15) and only one physio axes i.e., blood transplant rejection has a positive correlation with progression marker. This can also be meaning full since the blood transplant rejection axis has a very similar profile to the blood axis. However, changing the reference to HSC changes these monotonic behaviors. As an example, projecting ET samples on erythrocytes and macrophage axes of PhysioSpace does not show any more a high correlation with progression marker. As an example, the monotonic decreasing change of ET samples along the PC-based progression mark is shown in figure 38.

Results in PMF:

Projecting PMF samples on different physio axes show entirely different behavior than ET samples. When the reference of PhysioSpace is the blood tissue, interestingly 55 out of 66 show a high positive correlation (>+0.5) with the PC-based progression marker, and there is no axis with a high negative correlation. When shifting the reference of the coordinate system to HSC, only 5 axes are positively highly correlated with the progression marker, and 12 axes, including monocytes, leukocytes, macrophage, T cell, and blood, show a highly negative correlation with the progression marker (table 16). Interestingly PMF shows high positive scores on axes that are related to stemness, axes such as HSC, CD34+ blood, CD34+ blood thymus, embryonic stem cells, and multipotent progenitor stem cells; this could indicate that in PMF there might a differentiation block happen to prevent the formation of mature cells. Also, PMF samples show positive similarity scores to other progressive existing blood disorders such as plasma leukemia, multiple myeloma, and T cell-Acute lymphoid leukemia when the coordinate reference is blood (Supplementary figures).

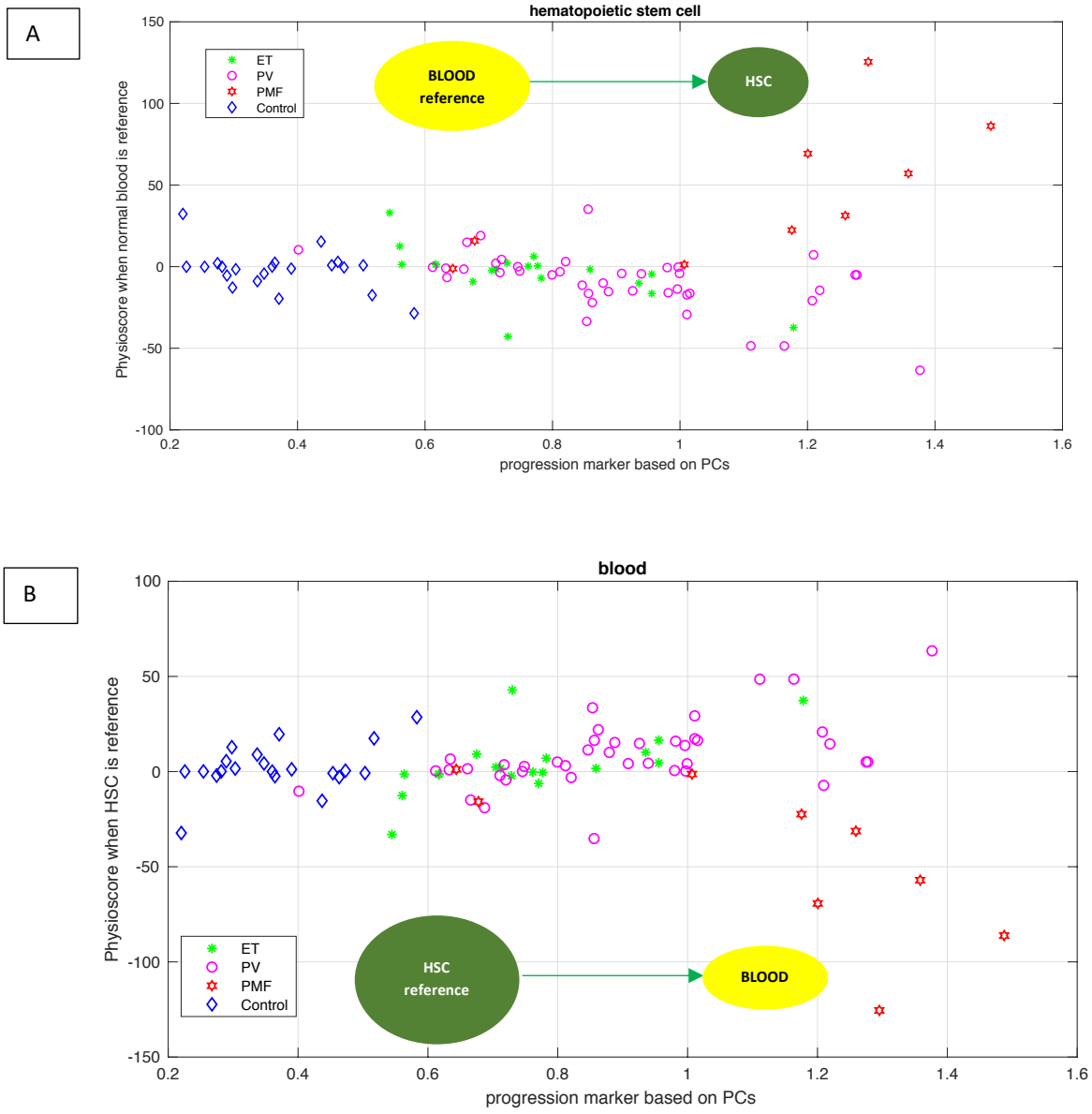


FIGURE 37: EFFECT OF CHOOSING HSCS OR WHOLE BLOOD AS THE REFERENCE OF PHYSIOSPACE COORDINATE SYSTEM

TABLE 14: LIST OF BLOOD-RELATED AXES IN PHYSIOSPACE

List of Blood-related axes in PhysioSpace		
• "AMO-1 myeloma" • "CD138+ plasma cell myeloma" • "FR4 multiple myeloma" • "KM4 multiple myeloma" • "NCI-H929 myeloma" • "NCU-MM1 multiple myeloma" • "SKMM1 multiple myeloma" • "multiple myeloma" • "rpmi-8226 myeloma" • "CCL-153 fetal pulmonary fibroblast" • "MRC5 fetal fibroblast " • "WI38 fetal fibroblast" • "adipose-derived adult stem cells" • "adipose-derived adult stem cells cultured" • "embryonic stem cell" • "hematopoietic stem cell" • "mesenchymal stem cell" • "multipotent adult progenitor cell" • "ccrf-cem T-ALL" • "CD34+ blood cell" • "CD34+ blood cell thymus" • "blood"	• "connective tissue fibroblast" • "embryonic lung fibroblast" • "embryonic skin fibroblast" • "fetal lung fibroblast" • "fibroblast Cockayne syndrome" • "primary fibroblast" • "progeria syndrome fibroblast" • "H1 mesenchymal precursor" • "H9 mesenchymal precursor" • "mesenchymal stem cell" • "CD138+ plasma cell" • "CD138+ plasma cell monoclonal gammopathy" • "Jurkat T lymphocyte" • "lymphocyte" • "lymphocyte asthma" • "B-cell lymphoma" • "B-cell lymphoma cell line" • "RJ2.2.5 Burkitt's lymphoma" • "SR lymphoma" • "blood transplant rejection" • "M70e hematopoietic"	• “CD138+ plasma cell myeloma" • "plasma cell" • "plasma-cell leukemia" • "B cell" • "B-cell lymphoma" • "B-cell lymphoma cell line" • "NC-NC lymphoblastoid B cell" • "MOLT4 T cell acute lymphoblastic leukemia" • "T cell" • "T cell diseased" • "ligament cell periodontitis" • "leukocyte" • "leukocyte multiple organ failure" • "erythrocyte" • "Su-dhl1 anaplastic large cell lymphoma" • "U937 lymphoma" • "ts anaplastic large cell lymphoma" • "monocyte" • "monocyte familial hypercholesterolemia" • "monocyte multiple organ failure" • "macrophage" • "I-6 embryonic stem cell”

TABLE 15: CORRELATION OF PHYSIOSCORES WITH PC-BASED PROGRESSION MARKER FOR ET SAMPLES

Physio axes	Correlation score When Blood is the reference	Correlation score When HSC is the reference
AMO-1 myeloma	-0.781803902	-0.539819112
CD138+ plasma cell myeloma	-0.251657012	0.461233534
FR4 multiple myeloma	-0.819352888	-0.787804869
KM4 multiple myeloma	-0.81797454	-0.701037815
NCI-H929 myeloma	-0.616090547	0.137685941
NCU-MM1 multiple myeloma	-0.754675768	-0.513781505
SKMM1 multiple myeloma	-0.867669576	-0.798005523
multiple myeloma	-0.349729845	0.212416001

rpmi-8226 myeloma	-0.490379206	-0.062079996
CCL-153 fetal pulmonary fibroblast	-0.770114694	-0.47095033
MRC5 fetal fibroblast	0.00405069	0.550763715
WI38 fetal fibroblast	-0.358571483	0.513907817
connective tissue fibroblast	-0.634156119	-0.357654483
embryonic lung fibroblast	-0.732860231	-0.420082604
embryonic skin fibroblast	-0.784320567	-0.421741637
fetal lung fibroblast	-0.664333664	-0.132560504
fibroblast Cockayne syndrome	-0.653519827	-0.416176479
primary fibroblast	-0.667262322	-0.082195426
progeria syndrome fibroblast	-0.639567385	-0.151838394
H1 mesenchymal precursor	-0.78104324	-0.564312979
H9 mesenchymal precursor	-0.499608619	0.41783427
mesenchymal stem cell	-0.611563571	0.129603801
CD138+ plasma cell	0.278534497	0.597028905
CD138+ plasma cell monoclonal gammopathy	0.248032472	0.535547764
CD138+ plasma cell myeloma	-0.251657012	0.461233534
plasma cell	0.2550428	0.581366732
plasma-cell leukemia	-0.679354359	0.23249751
B cell	-0.804178259	-0.661314614
B-cell lymphoma	-0.78349082	-0.59740792
B-cell lymphoma cell line	-0.553962572	0.277420731
NC-NC lymphoblastoid B cell	-0.810619351	-0.700880504
MOLT4 T cell acute lymphoblastic leukemia	-0.814033931	-0.776265336
T cell	-0.826759064	-0.60798817
T cell diseased	-0.787994349	-0.508831862
ligament cell periodontitis	-0.678382017	-0.415637554
leukocyte	-0.3008102	0.296489501
leukocyte multiple organ failure	0.194395948	0.54601085
erythrocyte	-0.756095401	-0.456278188
Jurkat T lymphocyte	-0.801220812	-0.756910605

lymphocyte	-0.798582974	-0.667732917
lymphocyte asthma	-0.794153076	-0.498566958
B-cell lymphoma	-0.78349082	-0.59740792
B-cell lymphoma cell line	-0.553962572	0.277420731
RJ2.2.5 Burkitt's lymphoma	-0.825984149	-0.74980378
SR lymphoma	-0.74430341	-0.606134024
Su-dhl1 anaplastic large cell lymphoma	-0.73874296	-0.56950041
U937 lymphoma	-0.750930579	-0.560217567
ts anaplastic large cell lymphoma	-0.751909337	-0.573297628
monocyte	-0.486548007	0.191229449
monocyte familial hypercholesterolemia	-0.091130364	0.661394251
monocyte multiple organ failure	-0.39818628	0.35772784
macrophage	-0.664955983	-0.415840271
l-6 embryonic stem cell	-0.389188939	0.259768713
adipose-derived adult stem cells	-0.21701788	0.389440234
adipose-derived adult stem cells cultured	-0.558281632	-0.311035441
embryonic stem cell	-0.762123509	-0.611900239
hematopoietic stem cell	-0.696329576	
mesenchymal stem cell	-0.611563571	0.129603801
multipotent adult progenitor cell	-0.624492901	-0.364590858
ccrf-cem T-ALL	-0.776004239	-0.809082972
CD34+ blood cell	-0.738439378	-0.575202456
CD34+ blood cell thymus	-0.655616392	-0.5395698
blood		0.696329576
blood transplant rejection	0.768438714	0.814367627
M70e hematopoietic	-0.518948807	0.252660856

TABLE 16: CORRELATION OF PHYSIOSCORES WITH PC-BASED PROGRESSION MARKER FOR PMF SAMPLES

Physio axes	Correlation score When Blood is the reference	Correlation score When HSC is the reference
AMO-1 myeloma	0.631439552	-0.046632742
CD138+ plasma cell myeloma	0.738904807	-0.398558488
FR4 multiple myeloma	0.638753301	0.323183953
KM4 multiple myeloma	0.613086607	0.211299425
NCI-H929 myeloma	0.751775984	0.128919702
NCU-MM1 multiple myeloma	0.632195746	0.005153807
SKMM1 multiple myeloma	0.450478666	-0.124354094
multiple myeloma	0.884811569	-0.244273097
rpmi-8226 myeloma	0.762491218	0.338657371
CCL-153 fetal pulmonary fibroblast	0.596619567	0.407287133
MRC5 fetal fibroblast	0.748622694	0.051249895
WI38 fetal fibroblast	0.770056833	0.266640508
connective tissue fibroblast	0.552461341	0.307423684
embryonic lung fibroblast	0.58647843	0.36360937
embryonic skin fibroblast	0.628238543	0.448757634
fetal lung fibroblast	0.629037856	0.316593254
fibroblast Cockayne syndrome	0.676257931	0.577672429
primary fibroblast	0.73357785	0.782784371
progeria syndrome fibroblast	0.687274082	0.679649083
H1 mesenchymal precursor	0.614169644	0.416054126
H9 mesenchymal precursor	0.69269454	0.464073327
mesenchymal stem cell	0.685345252	0.467937636
CD138+ plasma cell	0.609444876	-0.323150818
CD138+ plasma cell monoclonal gammopathy	0.605008169	-0.330132352
CD138+ plasma cell myeloma	0.738904807	-0.398558488
plasma cell	0.601071208	-0.313322333

plasma-cell leukemia	0.6522521	-0.53699681
B cell	0.517986456	-0.217749867
B-cell lymphoma	0.666533293	-0.525379735
B-cell lymphoma cell line	0.735617177	-0.091684258
NC-NC lymphoblastoid B cell	0.343643474	-0.539965464
MOLT4 T cell acute lymphoblastic leukemia	0.607090827	0.25205023
T cell	-0.348605706	-0.849668929
T cell diseased	-0.259600323	-0.876396476
ligament cell periodontitis	0.504119116	0.187216275
leukocyte	-0.19917417	-0.79395062
leukocyte multiple organ failure	0.532223209	-0.39968577
erythrocyte	0.563565715	-0.118413391
Jurkat T lymphocyte	0.589574734	0.20122718
lymphocyte	0.268627876	-0.483260742
lymphocyte asthma	-0.295779335	-0.900107656
B-cell lymphoma	0.666533293	-0.525379735
B-cell lymphoma cell line	0.735617177	-0.091684258
RJ2.2.5 Burkitt's lymphoma	0.711389246	0.4133354
SR lymphoma	0.547177659	-0.032946479
Su-dhl1 anaplastic large cell lymphoma	0.453255161	-0.106767438
U937 lymphoma	0.622614103	0.421863486
ts anaplastic large cell lymphoma	0.518557261	-0.018355323
monocyte	0.274737246	-0.655673985
monocyte familial hypercholesterolemia	0.671855483	-0.492506228
monocyte multiple organ failure	0.264644712	-0.587533042
macrophage	0.255722322	-0.600715513
I-6 embryonic stem cell	0.862836362	0.570038153
adipose-derived adult stem cells	0.688863741	-0.276548864
adipose-derived adult stem cells cultured	0.581361229	0.191614568
embryonic stem cell	0.722844098	0.674180456

hematopoietic stem cell	0.665199704	
mesenchymal stem cell	0.685345252	0.467937636
multipotent adult progenitor cell	0.625570624	0.37464983
ccrf-cem T-ALL	0.652365426	0.231326558
CD34+ blood cell	0.630112761	0.429016722
CD34+ blood cell thymus	0.434997042	-0.004522831
blood		-0.665199704
blood transplant rejection	0.431171112	-0.469983039
M70e hematopoietic	0.834224617	0.439203686
hematopoietic stem cell	0.665199704	

Results in PV

In comparison to ET and PMF that 71% and 83% of physio axes show monotonic changes over PC-based progression marker respectively, only 35 % of physio axes are correlated with PC-based progression marker (negative correlation coefficient <-0.5). This means that while PMF and ET have dominant behaviors, PV samples are acting somewhere between PMF and ET. This indicate that PV may be an intermediate stage between ET and PMF.

TABLE 17: CORRELATION OF PHYSIOSCORES WITH PC-BASED PROGRESSION MARKER FOR PV SAMPLES

Physio axes	Correlation score When Blood is the reference	Correlation score When HSC is the reference
AMO-1 myeloma	-0.546534011	-0.275679302
CD138+ plasma cell myeloma	-0.222773155	0.394080154
FR4 multiple myeloma	-0.606315708	-0.53830435
KM4 multiple myeloma	-0.570060868	-0.393567838
NCI-H929 myeloma	-0.538183411	0.032571614
NCU-MM1 multiple myeloma	-0.52231292	-0.259796772
SKMM1 multiple myeloma	-0.595362283	-0.548039281
multiple myeloma	-0.203994707	0.332073441
rpmi-8226 myeloma	-0.474119815	0.002588841
CCL-153 fetal pulmonary fibroblast	-0.481736276	-0.324675978
MRC5 fetal fibroblast	-0.030306588	0.439919952

WI38 fetal fibroblast	-0.243583731	0.39916167
connective tissue fibroblast	-0.417222525	-0.111159423
embryonic lung fibroblast	-0.447053559	-0.250084558
embryonic skin fibroblast	-0.485520823	-0.259795805
fetal lung fibroblast	-0.403542113	0.22071308
fibroblast Cockayne syndrome	-0.36540735	0.045721181
primary fibroblast	-0.350730756	0.239007056
progeria syndrome fibroblast	-0.333153406	0.294174413
H1 mesenchymal precursor	-0.448677689	-0.23367754
H9 mesenchymal precursor	-0.316225947	0.460544284
mesenchymal stem cell	-0.308510024	0.423224313
CD138+ plasma cell	0.217683318	0.474180195
CD138+ plasma cell monoclonal gammopathy	0.090103785	0.43453858
CD138+ plasma cell myeloma	-0.222773155	0.394080154
plasma cell	0.316302328	0.521263588
plasma-cell leukemia	-0.480782824	0.291395534
B cell	-0.591696769	-0.565494232
B-cell lymphoma	-0.519552442	-0.192811272
B-cell lymphoma cell line	-0.486963672	-0.014667809
NC-NC lymphoblastoid B cell	-0.57714521	-0.532277788
MOLT4 T cell acute lymphoblastic leukemia	-0.613149514	-0.647275629
T cell	-0.666817007	-0.680060996
T cell diseased	-0.611590064	-0.45552699
ligament cell periodontitis	-0.501329496	-0.334507937
leukocyte	0.073030721	0.39487933
leukocyte multiple organ failure	0.457613412	0.620402291
erythrocyte	-0.48263622	-0.323766916
Jurkat T lymphocyte	-0.580221359	-0.572020809
lymphocyte	-0.552163578	-0.464821785
lymphocyte asthma	-0.671746044	-0.638822635

B-cell lymphoma	-0.519552442	-0.192811272
B-cell lymphoma cell line	-0.486963672	-0.014667809
RJ2.2.5 Burkitt's lymphoma	-0.544007055	-0.367649163
SR lymphoma	-0.493847417	-0.385622591
Su-dhl1 anaplastic large cell lymphoma	-0.487824506	-0.410176284
U937 lymphoma	-0.499588636	-0.449129809
ts anaplastic large cell lymphoma	-0.502528415	-0.422717821
monocyte	-0.16158554	0.310216294
monocyte familial hypercholesterolemia	0.117811695	0.598743483
monocyte multiple organ failure	0.034483415	0.44080829
macrophage	-0.441805083	-0.238452227
I-6 embryonic stem cell	-0.285453956	0.173577032
adipose-derived adult stem cells	0.121920642	0.660740526
adipose-derived adult stem cells cultured	-0.288759003	0.163783282
embryonic stem cell	-0.530010846	-0.324963638
hematopoietic stem cell	-0.545326108	
mesenchymal stem cell	-0.308510024	0.423224313
multipotent adult progenitor cell	-0.350787706	-0.054473338
ccrf-cem T-ALL	-0.612094407	-0.628199135
CD34+ blood cell	-0.488415422	-0.400349807
CD34+ blood cell thymus	-0.484141656	-0.415666811
blood		0.545326108
blood transplant rejection	0.659517913	0.613922509
M70e hematopoietic	-0.492349671	-0.058227673
hematopoietic stem cell	-0.545326108	

5-1-4C: RESULTS OF PROJECTION OF IZKF INTO PHYSIOSPACE

When projecting IZKF data into PhysioSpace, we did not find any Physio axes which is significantly different between any of the MPN stages. The distribution of the P-value of an ANOVA test on 66 tested PhysioScores for both IZKF and GSE26049 datasets is shown in figure 39. We see that while there are significantly different physio axes between MPN stages in GSE26049 (P-value<0.01 and log10(P-value) <-2), there is no significantly different axes in IZKF dataset. This is because the IZKF data set was obtained from CD34+ purified cells, and

this means that they are not a mixture of cells. Therefore, different cell types and tissues are not present in them. This is an exciting finding supporting the idea that observed progression in the mixture of cells (figure 33) is a matter of shift of cell populations in blood tissue.

5-1-5: IMPROVEMENT OF GENE EXPRESSION BASED-CLASSIFICATION WITH PHYSIOSCORES

Our results in the previous section indicate that the projection of MPN gene expression in PhysioSpace could shed light on the different behavior of ET, PV, and MF. We also concluded that classification based on PCs:

- perfectly separates healthy samples from MPNs
- can classify PVs with a high accuracy
- but fails to distinguish ET and PMF.

Now, we are interested to see if PhysioSpace can help us to increase the classification accuracy. Therefore, based on the result of the first classifier, which is able to classify PV and control very well, we train the second classifier on the non-PV and non-control samples to predict the classes as either ET and MF. As we have 66 physio scores for each sample, high complexity leads to overfitting of the classifier; therefore, prior to the classification task, we do PCA on the PhysioScores. The question that the second classifier is designed to answer is: knowing that a sample is neither healthy nor PV, and having the PCs of PhysioScores, is the new sample an ET patient or an MF patient? The workflow of the model is shown in figure 40.

The cross-validation on the data resulted in an overall mean accuracy of 84% for training set and 78% for test set when training an SVM using PCs of PhysioScores. This classifier, in detail, does a perfect classification of ETs and predicts PMF with 67% accuracy. In order to investigate if the information content of PhysioScores helps to build a better classifier, we also did the same routine except that the inputs of the second classifier inputs are the results of the deconvolution algorithms for cell mixture analysis. The candidate for cell mixture analysis is the *CellMix*[188] toolbox *developed* by Gaujoux and colleagues. The complete technical details of using the *CellMix* is described in the appendix section A3. This package uses a list of genes defined by Abbas and colleagues[189], which contains immune specific genes identified from a compendium of microarray expression. The results of using the *CellMix* tool are the ratio of different immune system cell types such as different subtypes of T cells and different subtypes of B cells in the gene expression. To make it comparable with vectors of PhysioScores, we summed up all subtypes of a more general category to create one united ratio. For example, if there are 5 subtypes of T cells, we summed all of them to make the ratio of whole T cells. Therefore, in the end, we have the ratios of 7 cell types feeding the second classifier, including T cell, B cell, plasma cells, NK cells, monocytes, neutrophils, and dendritic cells. The workflow of using cell mixture analysis for the second classifier is shown in figure 41.

The second classifier with Cell mixture analysis results as input could also perfectly classify the ET, but it failed to classify the PMFs (55.5% accuracy) better than the classifier with PCs of Physio scores as input.

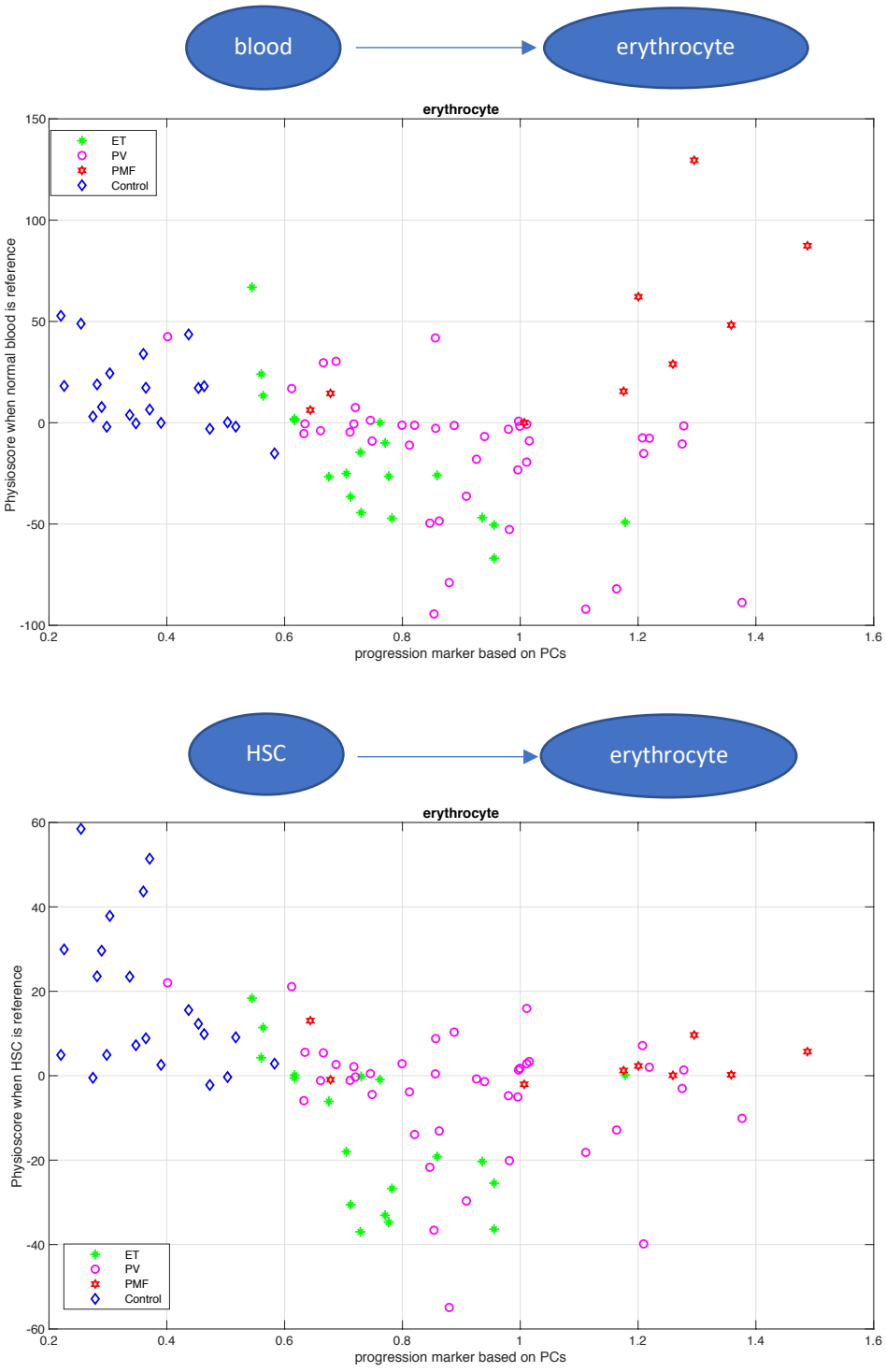


FIGURE 38: BEHAVIOR OF ERYTHROCYTE PHYSISCORES AGAINST PC-BASED PROGRESSION MARKER IN WHEN EITHER BLOOD IS THE REFERENCE (A) OR HSC IS THE REFERENCE (B) OF THE COORDINATE SYSTEM

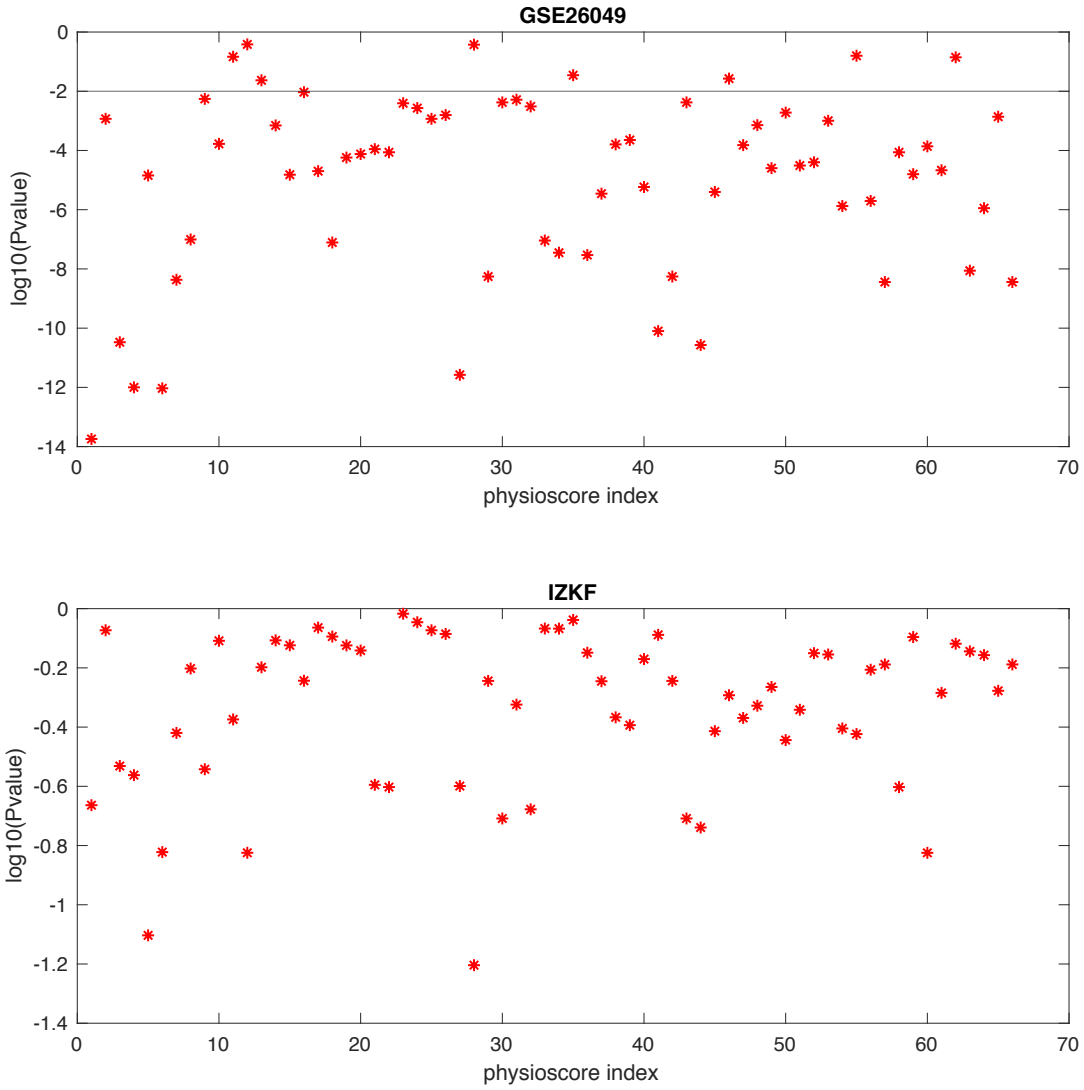


FIGURE 39: P-VALUE OF ANOVA TEST OF PHYSIO AXES BETWEEN DIFFERENT STAGES OF MPNS IN GSE26049 (UPPER PANEL) AND IZKF DATA SET (LOWER PANEL)

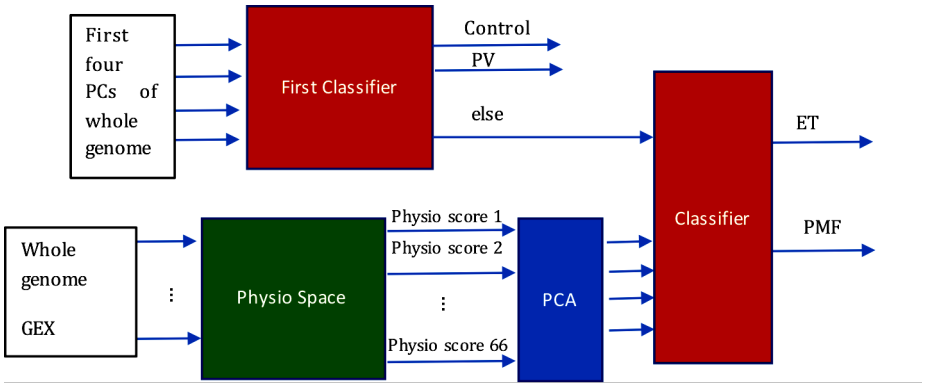


FIGURE 40: THE FIRST CLASSIFIER IS BASED ON PCS OF WHOLE GEX. THE SECOND CLASSIFIER IS ON THE PCS OF PHYSIOSPACE ANALYSIS OF GEX. THE NUMBER OF INPUT SAMPLES OF THE FIRST CLASSIFIER IS 91 (19 ET, 41 PV, 9PMF, 21 CONTROL). NUMBER OF SAMPLES FOR THE SECOND CLASSIFIER IS 28 (9PMF, 19 ETS)

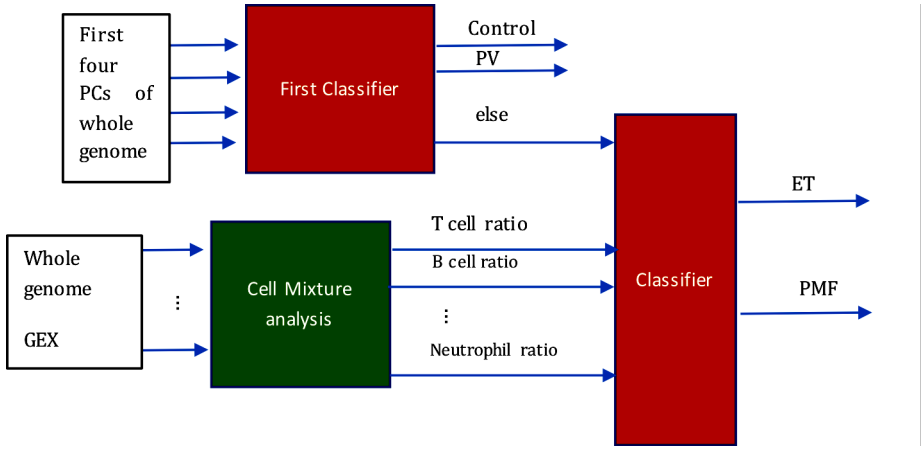


FIGURE 41: THE FIRST CLASSIFIER IS BASED ON PCS OF WHOLE GEX. THE SECOND CLASSIFIER IS ON THE RESULT OF THE CELL MIXTURE ANALYSIS OF GEX. THE NUMBER OF INPUT SAMPLES OF THE FIRST CLASSIFIER IS 91 (19 ET, 41 PV, 9PMF, 21 CONTROL). NUMBER OF SAMPLES FOR THE SECOND CLASSIFIER IS 28 (9PMF, 19 ETS)

5-2 HYBRID MODEL FOR CLASSIFICATION OF MPNS BASED ON CLINICAL VARIABLES.

Most of the time, in clinics, there is no access to gene expression arrays for better diagnosis, and diagnosis made based on cell counts. However, these diagnoses may not always be valid. In this section, we analyze the GSG-MPN dataset. In the GSG-MPN registry data, we have access to other kinds of data, for example, therapy information, age, medical checks such as if sonography is done or not and etc. These variables are interesting to study. At first glance, it may not be trivial that such data could carry information about disease development. But consider age as an example. Age is barely used in the diagnosis of the disease, but as we age, the immune system may lose its efficiency, and this may lead to some pathway impairment and a higher risk of cancer development. We can think the same way for other variables that they may relate to the disease in a non-trivial way. Different variables may connect together in a nonlinear way. This is interesting to study if these nonlinear relationships can be related to disease development. On the other hand, as our model is totally data-driven and is indeed black-box modeling, mixing all the variables in one pot and feeding them as input to the selected machine learning algorithms, raises the problem of curse of dimensionality and an exponential increase in the complexity of the model. This leads to extreme overfitting of the model on the training data so that extrapolating the model on the test set and new data is impossible. In 2002 Mogk and colleagues[122] showed that splitting one big black-box into sperate smaller sub-models reduces the number of inputs for each smaller sub-model and therefore reduces the complexity significantly. Therefore, we developed a hybrid model to test whether different variables excluding cell count carry information for a better classification. At first, we set up a classifier which is solely based on cell counts. This classifier's output is the probability that one patient belongs to each class. Based on this probability, we can calculate a misclassification score indicating how far is the predicted class from the original (diagnosed) labels in the registry. Secondly, we design a regression model based on other available variables, excluding cell counts, to predict the calculated misclassification error of the cell-count based classifier. Then having the predicted labels from the first classifier, we can use the predicted misclassification error to adjust the first predicted labels and improve the final classification accuracy. The reason while we split input space to cell count and non-cell count data is that, in the concept of hybrid modeling most of the time, we split the data space into

different regions depending availability of mechanistic models describing different parts of the whole system. But since here for MPN, we lack a mechanistic model such as Dingli model[121] for CML, and by knowing that MPNs are blood diseases leading to abnormal cells counts in different lineages, we try to substitute the non-existent mechanistic model for the cell population dynamics with a data-driven black-box model. The workflow of the hybrid model is shown in figure 42.

5-2-1: TRAINING CLASSIFIER BASED ON CELL COUNTS

In this section, we use data from the GSG-MPN registry. Data and preprocessing steps are later described in detail in the appendix (section A2-2). We have overall 572 patients in 5 different MPN stages: ET, PV, PMF, Post ET-MF, and Post PV-MF. The first classifier is designed to predict the correct label of each MPN patient based on the cell counts data. The cell count data contains the number of leukocytes, and the ratio of WBC cell types such as neutrophil, lymphocytes, monocyte, eosinophil, and basophil and also hemoglobin, hematocrit, and platelets. We tested different scenarios to pick the machinery with the highest classification accuracy:

- Selection of the best machine learning algorithm:
  We chose the best classifier based on the highest accuracy among NN, logistic regression, SVM and RF

- Selection of the best labeling:
  As we already mentioned, there are 5 categories of MPNs in GSG-MPN dataset: ET, PV, PMF, Post-ET-MF and Post-PV-MF. In this step, we tried to discover if the cell count data can classify MPNs truly in 5 separate categories or if Post-ET/PV MFs can be categorized together with MF as the last and third category. We also tested how the ordinal or categorical labeling can change the accuracy of the model. By ordinal labeling, it is meant that there is a spectrum of disease progression starting from ET as the least malignant version of MPN, continuing through PV, MF, and ending with secondary myelofibrosis as the most malignant version of MPN. By the categorical labeling, we do not consider any order of disease progression.

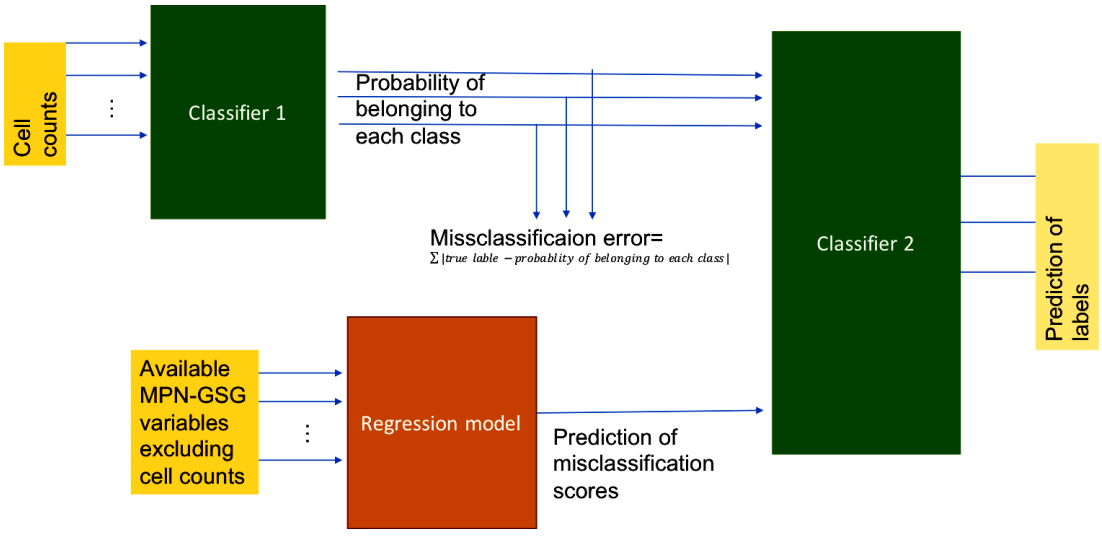


FIGURE 42: WORKFLOW OF THE HYBRID MODEL FOR MPN-CLASSIFICATION

So, for different labeling scenarios, we trained multiple classifiers and picked the algorithm and best labeling option. For example, considering the categorical labeling, we trained a logistic regression with five distinct categories. Interestingly, the classifier failed to diagnose any of the secondary myelofibrosis patients. The accuracy of the classifier for different categories is 78%, 80.6%, 52%, 0% and 0% for ET, PV, PMF, post-ET-MF and post-PV-MF respectively. When we take a closer look at the last two categories, 13 out of 19 (68.4%) of post-ET-MF patients are classified as PMF, and the rest are equally classified either as ET or PV (3 patients for each). In post-PV-MF, 9 of 17 patients (~53%) are classified as PV, 6 of them (35%) as PMF, and one of them as post-ET-MF. Poor accuracy for MF and null accuracy for POST ET/PV MF classification implies that the cell count data alone don't carry enough information for classification of these stages of MPN diseases. Based on these results, we designed a classifier with a new set of labels distributing patients in three major groups as ET, PV, and myelofibrosis (MF). In this case, all MF patients, including primary and secondary, are grouped together to form a group. This improves the overall classification accuracy from 67.8% to 72.25%. The final classifier based on the results, which are presented in the appendix (section A2-2, table A4), is selected to be a logistic regression model. The final outputs of this classifier are either ET, PV, or MF without any ordinal meaning with the following accuracies:

TABLE 18: CLASSIFICATION ACCURACY OF THE FIRST CLASSIFIER WITH MERGED CATEGORICAL LABELING

MPN type	Classifications accuracy
ET	76.67%
PV	75%
MF	53.84%
ETMF	78.95%
ETPV	58.82%

5-2-2: CALCULATION OF MISCLASSIFICATION SCORE

The outputs of the first classifier are initially the probabilities of belonging to each group of MPN; an example is shown in figure 43. On the other hand, we have the true labels (reported diagnosis in the registry) that can also bo shown as a probability of belong to each group. For example, suppose that the true labels for patient 1, patient 2 and patient n in figure 43 are ET, PV and MF respectively, true labels can be shown as the probability in figure 44.

	Probability of being ET	Probability of being PV	Probability of being MF
Patient 1	80%	15%	5%
Patient 2	30%	60%	10%
⋮	…	…	…
Patient n	40%	10%	50%

FIGURE 43: IN THE MATRIX ABOVE, EACH ROW REPRESENTS A PATIENT, AND EACH COLUMN THE CALCULATED PROBABILITY OF BELONGING TO A GROUP FROM THE CLASSIFIER

	Probability of being ET	Probability of being PV	Probability of being MF
Patient 1	100%	0%	0%
Patient 2	0%	100%	0%
⋮	…	…	…
Patient n	0%	0%	100%

FIGURE 44: IN THE MATRIX BELOW, EACH ROW REPRESENTS A PATIENT, AND EACH COLUMN THE PROBABILITY TRUE LABELS BASED ON REPORTED DIAGNOSIS:

Having these two matrices we can define a misclassification score (S) as:

$$S = \sum_{i=1}^{i=3} |reported\ probability\ for\ being\ class\ i - calculated\ probability\ for\ being\ class\ i| \quad \textbf{Eq. (11)}$$

The distribution of misclassification errors for each category is shown in figure 45. The lower the misclassification, the better the predicted classes resulted from our classifier. The maximum value of misclassification error is 2, and it is when the classifier fails to predict true labels. As we see in figure 45, when we do not merge the secondary myelofibrosis as MF and consider them as the classes of their own (figure 45-right), the misclassification is very high, while merging improves classification error in optimized labels (figure 45-left).

5-2-3: TRAINING A REGRESSION MODEL BASED ON CLINICAL DATA FOR ERROR PREDICTION OF THE FIRST CLASSIFIER

In this step, we want to test if other clinical variables, excluding cell counts as the main variables, could help for error adjustment of the MPN classification task. We assume that there is a non-linear cross-talk between different phenotypic layers, showing themselves as reported variables, which is non-trivial to extract. To test our hypothesis, we trained both NN and random forest on a refined selection of variables (with the highest availability in the registry). Here again, we tried different scenarios for model selections, such as a different combination of parameters as input and different network structures for NN. The complete process of choosing the best model and best set of variables are described in the appendix (sectionA2-3). Here, as the predicted value is continuous, we should train a regression model. The quality control is defined as:

$$error = \frac{y - y'}{0.5(y + y')} \qquad \textbf{Eq. (12)}$$

The denominator term is for normalization purposes. To summarize the errors in a large number of iterations of cross-validation, we calculated mean squared error as the final quality control. Based on our results in section A2-3, which are achieved by testing different selections of variables and models, a random forest using all the available variables as input precited the misclassification of cell-count-data-classifier the best. The complete list of input variables is shown in table A5 in the appendix (sectionA2-3). Error distribution of the final regression model in the test and training sets is shown in figure A14 in the appendix (sectionA2-3).

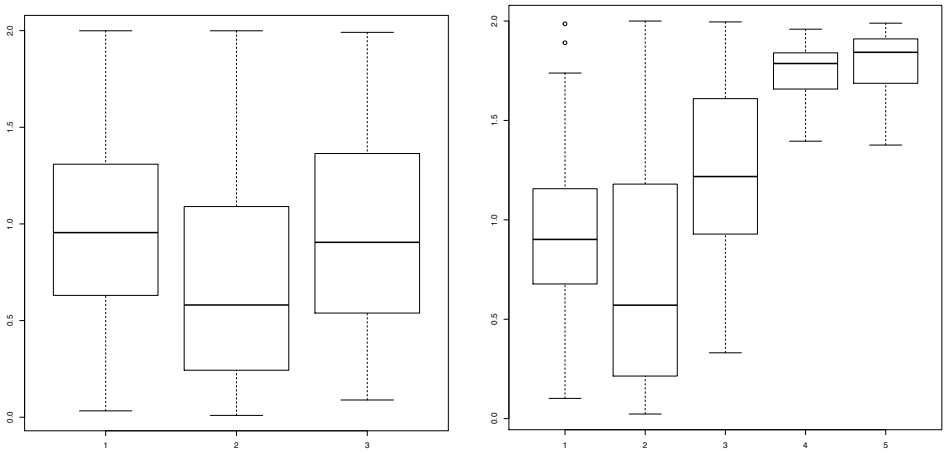


FIGURE 45: MISCLASSIFICATION SCORES ACROSS DIFFERENT DISEASE STAGES IN LEFT) MERGED LABELING, AND RIGHT) NON-MERGED LABELING

5-2-4: TRAINING A SECOND CLASSIFIER

So far, we trained two independent models with different variable sets as input. The first black model is a cell count based logistic regression classifier. Not every classifier is a perfect model with 100% accuracy. Therefore, we defined a misclassification error, showing how far are the predicted classes from the initially reported labels in the GSG-MPN registry. In the next step, we trained a random forest regression model on a set of other variables than cell counts, which are available in a reasonable number of samples from the registry to model/predict the misclassification error of the first classifier.

In this step, we are interested to see if the outputs of the sub-models in the previous steps can improve the MPN classifications. The third model is, therefore, a classifier with four input variables for each sample:

- probability of belonging to ET
- probability of belonging to PV
- probability of belonging to MF
- misclassification error

The first three inputs are the outputs of the first classifier, and the fourth input is the output of the random forest regression model.

We tried regularized neural network, logistic regression, and random forest as the candidate machine learning algorithms for the third black box. The complete description of model training is brought in the appendix (A2-4). The best accuracy is achieved by using a random forest model as the final classifier. Final results show improvement of classification accuracy in both training and test sets:

TABLE 19: COMPARISON OF CLASSIFICATION ACCURACY BETWEEN CELL COUNT BASED CLASSIFIER AND HYBRID CLASSIFIER

	Accuracy in train	Accuracy in test
Only cell count based classifier	72.25%	69.3%
Hybrid classifier	82.3%	83.11%

5-3 COMPARISON OF THE HYBRID MODEL PERFORMANCE WITH OTHER MODELS:

5-3-1 A MODEL ON THE WHOLE FEATURE SPACE

In order to check if this hybrid modeling outperforms other modeling strategies, we calculated classification accuracy, when all the variables are used straightforward for the class prediction. Here we apply logistic regression, random forest, and regularized neural network on all variables, including both the cells counts and the non-cell count data, as input to predict each sample's class (figure 46).

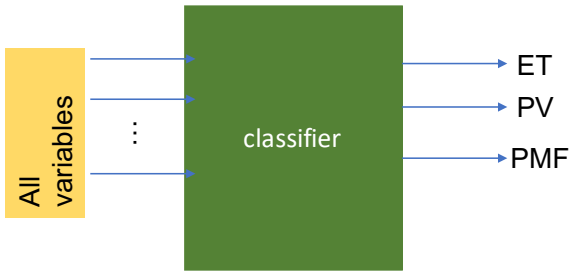


FIGURE 46: ALTERNATIVE CLASSIFIER WITH ALL VARIABLES AS INPUT

TABLE 20: ACCURACY OF DIFFERENT CLASSIFICATION ALGORITHM ON WHOLE INPUT SPACE

	Accuracy in train	Accuracy in test
Logistic regression	87.5%	72%
Random forest	78%	79%
Neural network	35%	36%

When using all the variables as input, we see that logistic regression, although it gives a better accuracy in the test set when compared to the cell count-based classifier, but it suffers from overfitting with a significantly higher training accuracy. In case of applying random forest on the whole variable space, although we have high accuracy in the training set and out of bag accuracy is also very high (95-100%), but when applying the generated forest on the new test set, we observe a much lower accuracy of the test set (68-76%). Regularization of random forest, although improves overfitting problem, the maximum test set accuracy is 79%, while our hybrid model could predict test sample's classes with an accuracy of 83.11%.

5-3-2 CLASSIFIER MODELS ON CLUSTERS OF NON- CELL COUNT DATA

We also tested another method to compare our hybrid model with. In this case, we run the clustering techniques on non-cell count data and trained logistic regression classifiers on the cell count data within each cluster (figure 47).

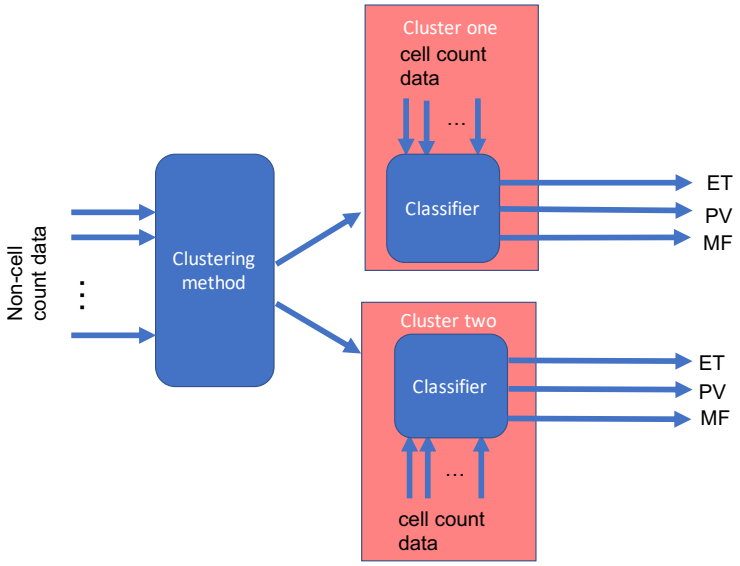


FIGURE 47: ALTERNATIVE CLASSIFIER ON CLUSTERS OF NON-CELL COUNT DATA

First, we performed hierarchical clustering, and second, we performed K-means clustering. For hierarchical clustering, we used *the hclust*[190] function from the *stats* package of R for the Euclidian distance matrix of the non-cell count data (figure 48). We then considered two clusters with 363 and 87 samples respectively in each. For K-means clustering, we used *the kmeans*[191] function of stats package in R and picked n=2 as the optimal number of clusters based on silhouette[192] width (figure 49). This resulted in the division of 359 and 91samples in different clusters. Based on our previous results, that logistic regression gives the best classification accuracy on the cell count data, we ran logistic regression classification (table 21) on resulted clusters. Accuracy of the classifiers in the training and test sets after are reported in table 21.

TABLE 21: CLASSIFICATION ACCURACY OF SAMPLES WITHIN DIFFERENT CLUSTERS OF NON-CELL COUNT DATA

		Training accuracy	Test accuracy
Hierarchical clustering	Cluster 1 (363 samples)	72%	68%
	Cluster 2 (87 samples)	75%	63%
K-mean clustering	Cluster 1 (359 samples)	72%	69%
	Cluster 2 (91 samples)	79%	68%

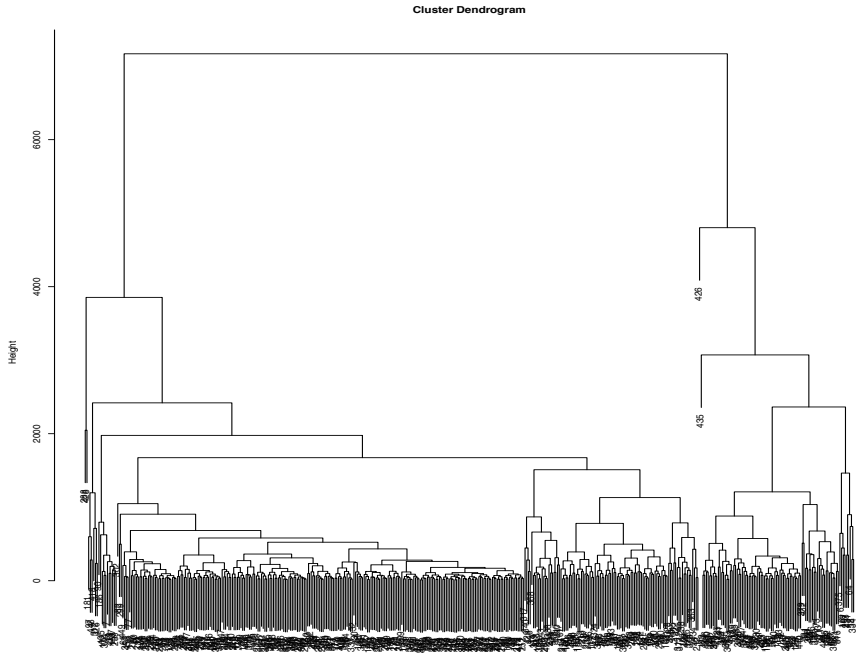


FIGURE 48: DENDROGRAM OF HIERARCHICAL CLUSTERING ON NON-CELL COUNT DATA

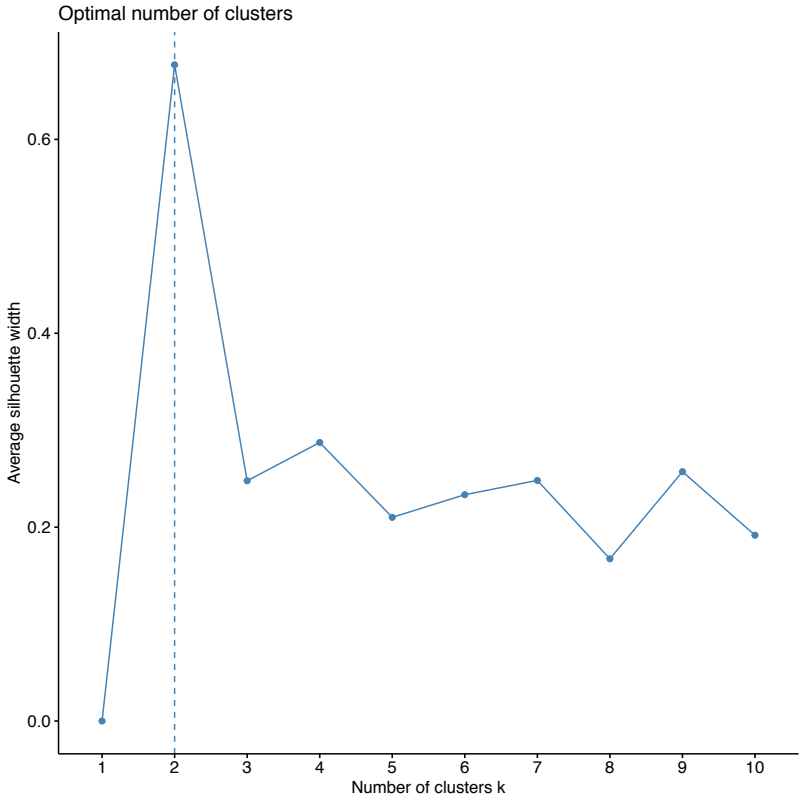


FIGURE 49: SILHOUETTE WITH FOR THE OPTIMAL NUMBER OF CLUSTERS.

5-3 CONCLUSION

In this chapter, we investigated the Philadelphia negative MPNs at two levels of the gene expression data and the clinical data.

At the gene expression level, we showed that the first principal components of the gene expression data space could reveal a progression trend from control to PMF via ET/PV. Therefore, we developed a PC-based biomarker, which is linearly increasing along with the MPNs evolution. We also investigated the association of gene expression profile of MPN patients with physiologically-related axes that are provided through PhysioSpace. Projection of Patients' gene expression data on PhysioSpace, revealed some interesting patterns inside different stages of MPNs. For example, in PMF patients, we observed higher PhysioScores in less differentiated cell type axes. Or we observed that along with the PC-based progression biomarker, most of the physio axes experience a monotonic change in ET patients. We also observed that in contrast to ET and PMF, we don't observe any dominant behaviors of Physio axes in PV patients. We further showed that these PhysioScores contain information that can improve the PC-based classifier's accuracy.

At the clinical data level, we developed a hybrid model consisting of the black box sub-models on subspaces of the original feature space. These subspaces are the cell count data the non-cell count data. The first black box sub-model is a classifier that is trained only on the cell count subspace of feature space. We tried different models such as NNs, SVMs, RFs, and logistic regression models for the purpose of classification. We found that a logistic regression model provides the best classification accuracy. The second sub-model is trained on the non-cell count data such that it models the classification errors of the first sub-model. After training both sub-models, we integrated them together for a better classification result. Based on our results, this hybrid modeling approach, not only can reduce overfitting problem of data-driven models arising because of the curse of dimensionality, but also can improve prediction accuracy.

When we compareed final gene expression and clinical data analyses, we can see that the hybrid clinical data-based classifier outperforms any gene expression-based classifier by the accuracy of 83% against 78% (table 22), respectively. This could state that data in the clinical level, which include cell count reports, therapy information, general health status, demographic information etc. are capable of distinguishing different MPN stages even more accurate than genotypic characteristics.

TABLE 22: SUMMARY OF ACCURACY RESULTS FROM DIFFERENT MODELING STRATEGY

Modeling strategy	Accuracy
GEX PC-based classifier	77%
GEX PC-based classifier coupled with PhysioScores	78%
Cell count based classifier	69%
Hybrid classifier	83%
Classifier on Hierarchical clusters	63-68%
Classifier on K-mean clusters	68%

CHAPTER 6: ENTROPY ANALYSIS

6-1 MOTIVATION

In chapter 4, we used entropy as an alignment marker for integrating patient-derived biomarker and the biomarker, which was based on Dingli's population dynamics model of CML evolution[125]. We showed with our simulations that entropy behaves reversely in gene expression level and cell population level during the disease progression of CP-CML. As the disease progresses, CML cells population exponentially explodes, and at the transition point, they take over the healthy cell population. Therefore, at this transition point, we observe the highest heterogeneity at cell population levels and, respectively the highest value (maximum) of Shannon entropy of the cell population. Our simulation also showed that the highest entropy at the cell population level is simultaneously happening at minimum entropy of gene expression level (figure 20). These results, however, are solely based on our simulation and implementation of Dingli's model of cell population growth in chronic CML. Here we tried to examine our assumption that maximum entropy at the cell population level occurs on minimum entropy at gene expression level as the disease progresses with the help of cell count data provided in the GSG-MPN dataset.

6-2 METHOD

We calculated Shannon entropy based on equation 7 at both of the gene expression level (GSE26049) and the cell population level (GSG-MPN data). For clarification, it is worth repeating that these are two different datasets with different patient cohorts. In equation 7, we need a Probability Distribution Function (PDF), which can be estimated by calculating the histogram function of each sample on the variables. Prior to calculating the histogram, we performed variable wise Z-score normalization, both for the cell count data and the gene expression data. Then we applied the *hist* function of MATLAB. To use histogram function instead of PDF, it is important to compare the histogram of all the samples in the same interval and in common bins. Otherwise, MATLAB histogram routine automatically produces bins based on the distribution of variables for each individual sample separately and without considering the whole dataset and other samples. Suppose in an experiment, all the gene expression values of patients' cohort cover the range of [-4,4], and patients are either in phase A of the disease or phase B. One patient in phase A of the disease has a distribution of gene expression in [-2,0]. Another patient in phase B has the same form of the distribution of gene expression but in the interval of [-1,3]. When we want to compare the frequency of each sample's gene expression, we have to estimate the PDFs over the entire input space, i.e. [-4, 4]. To calculate entropy with histogram we have to set the same bins for all samples. To do that for both datasets we do the following steps: find the minimum and maximum value in the whole dataset, then consider a number of bins *n*. Therefore, each smaller interval is 1/n of the complete interval and is the same for all sample. Now, we can assess the frequency of each smaller interval for each patient and calculate the Shannon entropy.

6-3 RESULT

We can plot the gene expression entropy against the PC-based gene expression marker (figure 50). But for the cell count data, as we do not have a continuous progression marker, it suffices to boxplot entropy against the disease stages. In figure 51, the upper panel shows the box plot of the entropy at the cell count level, which is calculated form the GSG-MPN dataset.

The lower panel shows the box plot of entropy at the gene expression level, which is calculated using GSE26049. As it is indicated in figure 51, the lowest median of gene expression entropy distributions belongs to PV, and interestingly, the highest median of cell count entropy is also occurring in PV. We observe a reverse behavior of entropy dynamics across MPN stages when compared across different levels of gene expression and cell count. This means that while the entropy (and respectively, the heterogeneity) of cell counts increases and reaches its maximum at the PV stage, at the gene expression level, the lowest entropy (and respectively the highest heterogeneity) happens again during PV. We can only compare ET-PV-PMF on both levels of data. That is because, in the patients' cohort belonging to gene expression data, we do not have any secondary myelofibrosis samples, and in the cell count level data, there is no healthy sample available.

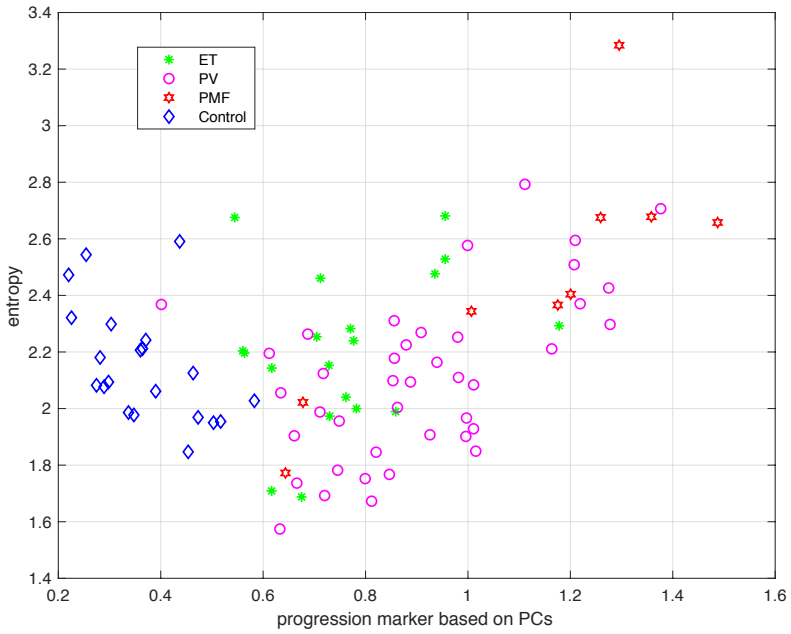
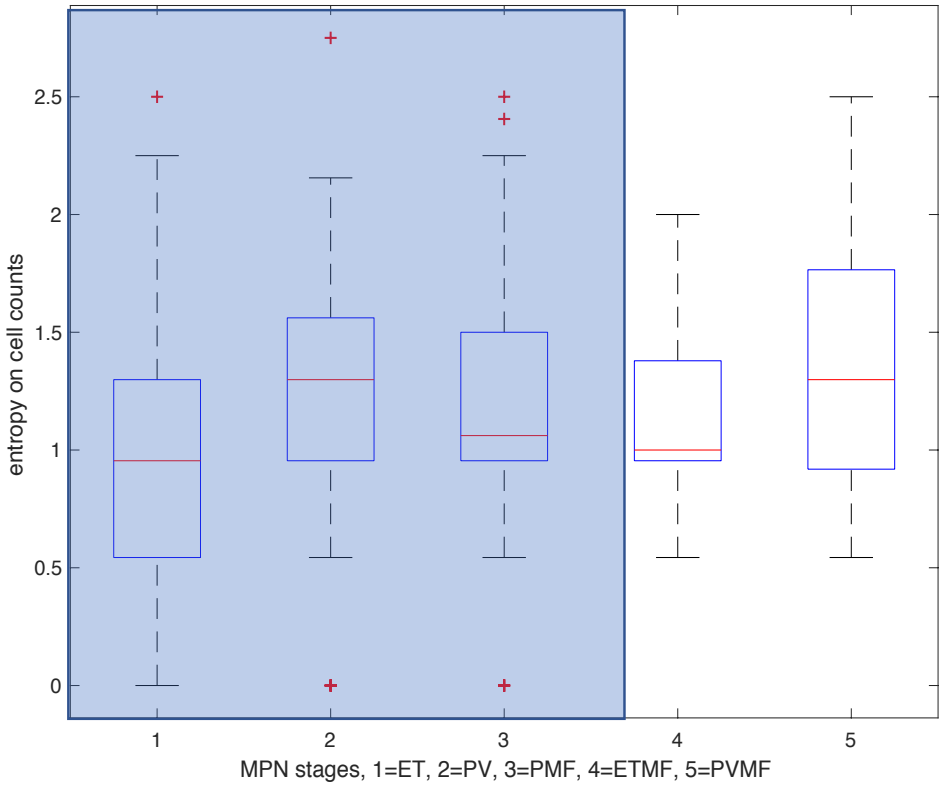


FIGURE 50: ENTROPY OF GENE EXPRESSION OF EACH MPN SAMPLES FROM GSE26049 AGAINST PC-BASED PROGRESSION MARKER

5-4 CONCLUSION

These results support our assumption that entropy at the gene expression level and the cell count levels behave reversely. This finding is interesting in the detection of the critical transition point, since finding the extremum of entropy in one level of data can help us to assess another level's extremum. However, to show that these extrema are happening at the same time in the disease progression time course, we need the cell count information and the whole blood gene expression data of the same sample cohort, which are unfortunately not available.



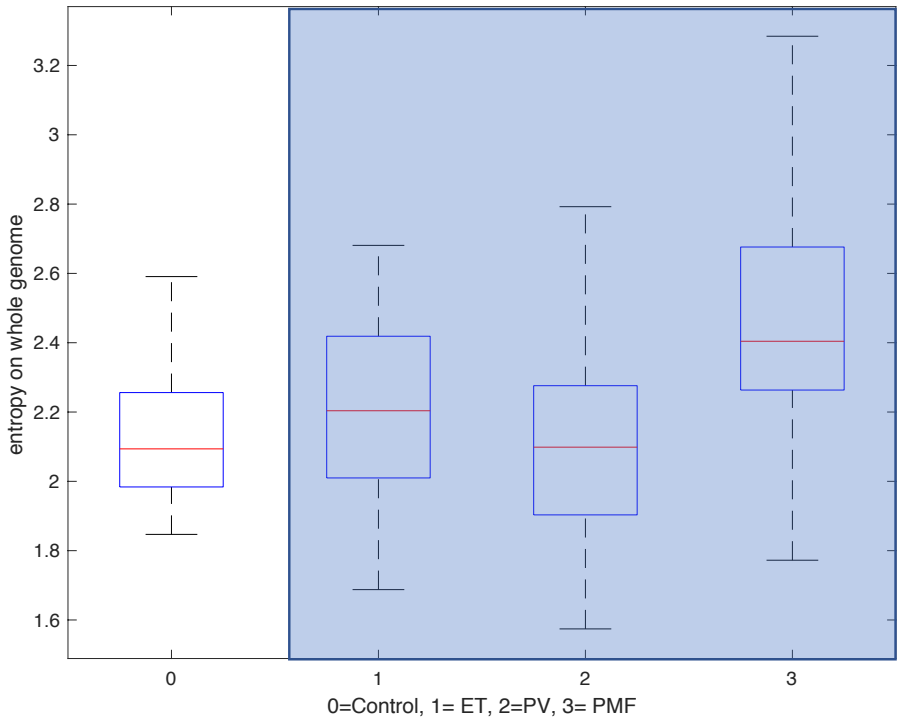


FIGURE 51: COMPARISON OF BOXPLOT OF SHANNON ENTROPY OF DIFFERENT MPN STAGES IN CELL COUNT LEVEL (A) AND GENE EXPRESSION LEVEL (B)

CHAPTER 7: CONCLUSIONS

Here in our research, we tried to model chronic disease progression by considering cancer as a multistep process. By having this insight into the multi-level complexity of cancers, we bring the urge of a multi-level modeling approach in focus. As already mentioned in the introduction and through the thesis, in the field of disease modeling, we not only need an average patient model of the disease progression, but we also need to integrate individual risk factors for developing patient-specific models[1]. The solution we presented in our research is hybrid modeling. We developed models at different levels, such as the cell populations and the gene expression, and integrated these models for disease progression tracking and more accurate diagnostic purposes. Specifically, we focus on disease modeling in myeloproliferative neoplasms. We designed our study in two different panels based genetic basis of MPNs: Philadelphia positive (CML) and Philadelphia negative MPNs (ET, PV, and PMF).

In the first panel and for CML, our hybrid models integrate a population-dynamic-based mechanistic model with gene expression data. The population dynamics model provides a simulation of how blood cells population in an average patient change in chronic CML. On the other hand, gene expression profiles are representing the genotypic changes of each individual sample. We invented the biomarkers that are achievable and correlated on both sides. With the help of the systems theory concepts and entropy, we could integrate these biomarkers, which are derived from heterogeneous sources of data. This hybrid approach helps to track disease evolutionary time in a quantitative manner, which is important for therapy management and prevention of entering the risk zone.

For the development of the hybrid model for Philadelphia negative MPNs, there is no available population-dynamic-based model. However, we have access to the GSG-MPN dataset, which includes cell count reports of MPN patients. Based on the cell count data, we set up the first sub-model of our hybrid model for the classification of different MPNs. The first sub-model is considered as the average patient model in analogy to our CML hybrid modeling approach. In the second sub-model, we utilized data such as age, gender, therapy information, patient history, etc. to individualize the average patient model for the purpose of more accurate classification. In fact, the second sub-model is used to compensate for the classification error rising from the first sub-model. The overall model, therefore, is a patient-specific model and helpful for better diagnostic purposes and personalized therapy management. Our results support that hybrid model not only can increase the classification accuracy but also is another example that hybrid modeling can be a solution to the overfitting problems of black-box modeling arising due to the curse of dimensionality. Also, in comparison to gene expression-based classifiers, our hybrid model on clinical laboratory data could provide better classification accuracy. This is an important finding because most of the time in clinics, clinical data is the only assessable data type, and there is no access to gene expression data. This model can be extended and improved by integrating more data types at more levels and steps of the complex system of disease progression. For example, gene expression data can add more insight to disease evolution from genomic alteration sight of the view. So, we recommend the assessment of different types of data from the same patients' cohort for the improvement of our model.

In the CML modeling part, we used entropy as an alignment marker for the integration of our two independent biomarkers. One of our hypotheses for using entropy as an alignment marker was its reverse behavior in two different levels of the cell population data and the gene expression data. This hypothesis was based on our simulation of Dingli's population-dynamic-

based model [125] for the CML evolution. Based on our simulation, we could see the minimum entropy in gene expression, and maximum entropy in cell population are aligned and indicative of the highest heterogeneous state in chronic CML. We could observe this reverse behavior of entropy again by calculating entropy on real gene expression data and real cell count data from two different cohorts of Philadelphia negative MPN patients. This is an important observation since this supports the idea that assessment of entropy extremum, which is accounted as a warning sign of critical transition [145], in one data level equals the assessment of entropy extremum in the other data level. However, to show it concisely, we need a gene expression and cell count data sets from the same cohort of patients.

In summary, we showed here that hybrid modeling in the sense of the combination of mechanistic modeling and data-driven modeling from the integration of heterogeneous sources of data could compensate the drawback of solely mechanistic and solely data-driven models and be helpful in the field of precision medicine where there is a need for patient-specific models.

APPENDIX: SUPPLEMENTARY METHODS AND MATERIALS

In this chapter, we try to provide the full detailed description of methods used for producing results of previous chapters. We also put some of the codes used in model selection for the purpose of reproducing results.

A1: DATA PREPROCESSING FOR GENE EXPRESSION ANALYSIS

For further analysis of choosing the right modeling algorithm for gene expression analysis, we should do some preprocessing such as dimension reduction and normalization. The dataset we used in this section is GSE26049. The array used for this dataset is the Affymetrix Human Genome U133 plus2.0. In their experiment Skov and Hasselbach[184] purified RNA from whole blood, amplified to biotin-labeled aRNA and hybridized to microarray chips. There are 21 healthy control samples, 19 ET patient samples, 41 PV patient samples and 9 PMF patient samples. The number of genes in the dataset is 17275. Data is downloaded via *Biobase*[193] and *GEOquery*[194] packages of R studio. Prior to run the principal component analysis, gene-wise Z-score[195] normalization is done so that all genes across all the samples have the mean of zero and standard deviation of one. To apply PCA we used SVD routine of MATLAB.[174]

The process of model selection is as follows:

- First, we do PCA for dimension reduction. Calculated PCs are called new features.
- Second, we choose a subset of new features.
- Third, we randomly split the data for cross-validation for 100 times, and train the candidate model with the selected feature subsets on the training set and evaluate on the test set.
- Fourth, we calculate the average accuracy of the model in the training and the test set.
- We repeat the second to fourth steps for different subsets of PCs.
- The model with the highest accuracy in the training and the test set is chosen as the final model

Candidate models are NNs, logistic regression, SVM and random forests.

A1-1: USING NEURAL NETWORK FOR CLASSIFICATION OF MPN FROM GENE EXPRESSION

The goal of applying NN is to classify MPNs based on gene expression profiles of patients. We aim to extract the potent trend of progression among MPNs and since in gene expression arrays we have thousands of dimensions, having an intuition of overall behavior of genes network in higher dimensions is not trivial. So, we need to significantly reduce the number of the data dimensions. For this purpose, we have applied PCA. PCA results although give us some information that gene expression array may be capable of indicating a trend of progression, but this information is not enough for having a descent biomarker capable of perfectly classifying MPNs with high accuracy (figure 33, table 11). Therefore, to help our restricted human capacity to perceive higher dimensions, we are obliged to take help from advanced pattern recognition patterns. With this hope that NN as a recent and powerful machine learning algorithm can help us for perceiving complex and possibly nonlinear alteration of genes pool in different stages of MPNS, we used NN to classify and reveal the hidden pattern. NNs are being translated in different programming languages such as Python,

R, MATLAB, etc. Here in this research, we used R language. As most of the computations in our research is done either in MATLAB or R studio, we looked for available and comfortable packages. One of the available NN packages in Rstudio is neuralnet[196], developed by Fritsch and colleagues. This package provides training deep NNs and the prediction of output for new data with the trained NN. This package also makes it possible to plot a trained network with all nodes and edges and corresponding weights. However, to use this package, we should be careful with special formatting of the in/output data.

A1-1A: DATA PREPROCESSING AND FORMATTING FOR USING NEURALNET PACKAGE:

In order to use neural network, we first ran PCA on normalized gene expression and then chose the first 10 PCs to train a NN. *neuralnet* fits a NN with backpropagation[197] algorithm. As already mentioned, in order to use *neuralnet* package, the data should have the special format. That is because, *neuralnet* package does not accept a simple labeling (classes of MPNs) of the data, such that every sample is given a category as 1,2,3 or 4 in a vector from. Instead we have to make a matrix for feeding into NN as labels in the form of "*one hot encoding*" or *dummy* encoding. That means we have to reform the class vector as a $\boldsymbol{n} \times \boldsymbol{m}$ matrix, where n is the number of samples and m is the number of classes (here MPN stages). If a sample belongs to a specific category or not, is shown with one or zero respectively. So, for example, if we have 7 patients in 3 categories coded by 1, 2 and 3 [1,1,3,2,3,2,2], the respective one hot encoded labeling matrix will be like this:

Patients	Category 1	Category 2	Category 3
Patient one	1	0	0
Patient two	1	0	0
Patient three	0	0	1
Patient four	0	1	0
Patient five	0	0	1
Patient six	0	1	0
Patient seven	0	1	0

A1-1B: APPLYING NEURAL NETWORK FOR CLASSIFICATION ON PCS

Now that we have the labeling matrix as output and the first 10 PCs as input, we can train a NN. Fortunately for using *neuralnet* package, we don't require other libraries. But for plotting the result we need the library of ggplot[198]:

The source code for starting *neuralnet* Package in R studio environment is:

```
>require(ggplot2)

>library (neuralnet)
```

The main function of this package for training a NN is *neuralnet (…)*. As inputs, we should provide data in *data.frame* format (both of the input and the output data should be in the same data structure), a formula which tells what are the inputs and outputs, here PCs as input and labels in form of dummy variables as output. Another input term is the architecture of neural networks, which is given as a vector of the number of nodes in each layer. For example, a vector [5,10,5] means that there are three hidden layers with 5 nodes in the first hidden layer, 10 nodes in the second hidden layer and 5 nodes in the third hidden layer. We also have to

set the action function as an input of the *neuralnet* (…). As we are interested in the classification task and not regression we should set out *act.fct* (action function) as *logistic* and not *linear.* After setting all the inputs of the function right, we are ready to train our NN. Arbitrarily I set the architecture as three hidden layers with 5 nodes in the first hidden layer, 10 nodes in the second hidden layer and 5 nodes in the third hidden layer. I split data randomly in 70% training and 30% test set and ran the neural network for 100 times. I did the same routine every 100 times for a different number of inputs, meaning a different number of PCs. Starting from only 2 PCs to 10 PCs. Each time the prediction accuracy of the labels in both of the training and the test set is calculated as the quality measure of the network. As a result, for each set of the input PCs, I have 100 neural networks and their corresponding accuracies, which I summarized as mean. The results are shown as the accuracy of test and training in figure A1.

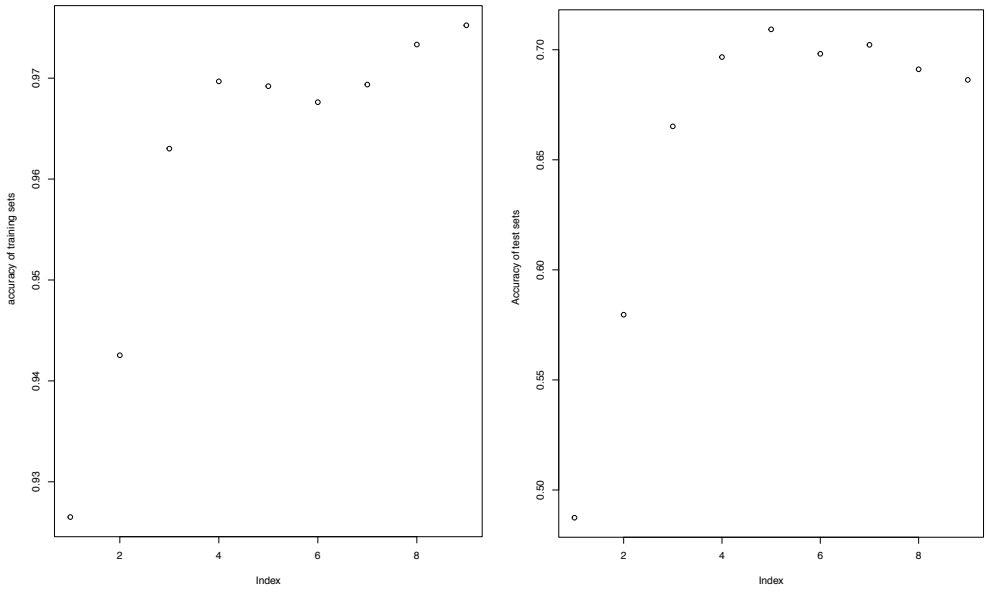


Figure A1) mean accuracy of trained NN with different number of PCs as input in training set (left) and test set (right)

The index on the x-axis in both sides of figure A1 indicates the number of PCs used to train the Neural Network. And the y-axis shows the calculated mean accuracy over 100 times of cross-validation for each number of the input PCs. As you can see the Network accuracy of the training set is quite high (93%-99%). On the other hand, the average accuracy is notably lower for test set when low number of PCs are fed to the system as inputs. Increasing the number of PCs as inputs increase classification accuracy in the training set. We can see that highest accuracy (almost 71%) in the test set however is achieved when we use 5 PCs as input data. Using more than 5 PCs leads to more complexity of the network, and therefore worse accuracy in the test set. But in a general conclusion, with a comparison between accuracies in the training and the test set, we observe the *overfitting* problem of NN for our dataset. Anyhow the results could mean that the five first PCs is the best set of input variables to classify the test sets labels with a NN. It should be also noted that the highest accuracy in the test set is 0.71 which is not considered to be perfect! In order to see if NN can perform better compared to simpler types of pattern recognition machinery, we also did the same routine with logistic regression.

A1-2: USING LOGISTIC REGRESSION FOR CLASSIFICATION OF MPN FROM GENE EXPRESSION

I performed multinomial logistic regression, with the help of two packages. First *nnet*[199], and second *glment*[200].

A1-2-1: MULTINOMIAL LOGISTIC REGRESSION WITH NNET PACKAGE.

The multinomial logistic regression is an extension of the logistic regression. It is used when the outcome involves more than two classes, which makes it a perfect choice for our study since we have more than one stage of MPN in our gene expression data.

A1-2-1A: DATA PREPROCESSING AND FORMATTING

In order to use logistic regression routine in the *nnet* package, we first ran PCA on normalized gene expression and then chose the first 10 PCs to train a logistic regression model. Here we don't need any special encoding of labels as we needed for the *neuralnet* (…) function in *neuralnet* package. A vector which codes with numbers different stages of MPN is enough.

A1-2-1B: APPLYING LOGISTIC REGRESSION (NNET) FOR CLASSIFICATION ON PCS

For running the logistic regression algorithm and training the logistic regression model, we used function is *multinom*, and to be able to apply this function we require other libraries extra to the main *nnet* package which are the followings:

```
>require(foreign)

>require(nnet)

>require(ggplot2)

>require(reshape2)
```

Multinom (…) function has multiple inputs, first the data which needs to be in *data.frame* format and then a formula which describes the in/outputs of the model. Inputs, here again, are PCs, for which we change each time the number of PCs in order to observe the effect of increasing dimensionalities or model complexity. Outputs are the labels describing the stage of disease for each patient. Again, I split the data randomly in 70% training and 30% test set to make it possible to compare with the results of NN. This splitting is done 100 times randomly. So overall, we have 100 separately trained model for each training set. This procedure is repeated for a different set of inputs i.e. a different number of PCs (starting from only with 2 PCs to 10 PCs). Each time the accuracy of prediction of labels in both training and test sets is calculated. As results, we have for each set of input PCs, 100 models and their corresponding accuracies, which again is summarized in the mean. The results are shown as the accuracy of the test and training sets in figure A2.

The index in both sides of figure A2 indicates the number of PCs used to train the logistic regression model starting from one to 10. As you can see the model accuracy in the training set is quite lower compared with the result of Neural network. Unlike NN, the accuracy in the test set is not much different from the accuracy in the training set. Similar to NN, increasing the number of PCs as the input increases the classification accuracy in the training set.

However, we can see that the highest accuracy (almost 69.7%) in the test set is achieved when we use 4 PCs in input. Feeding more than 4 PCs as the input increases the complexity of the model and decreases the model accuracy in the test set. This means that the four first PCs is the best set of input variables to classify the test sets labels. However, the highest accuracy is 69.7% which is not considered to be perfect!

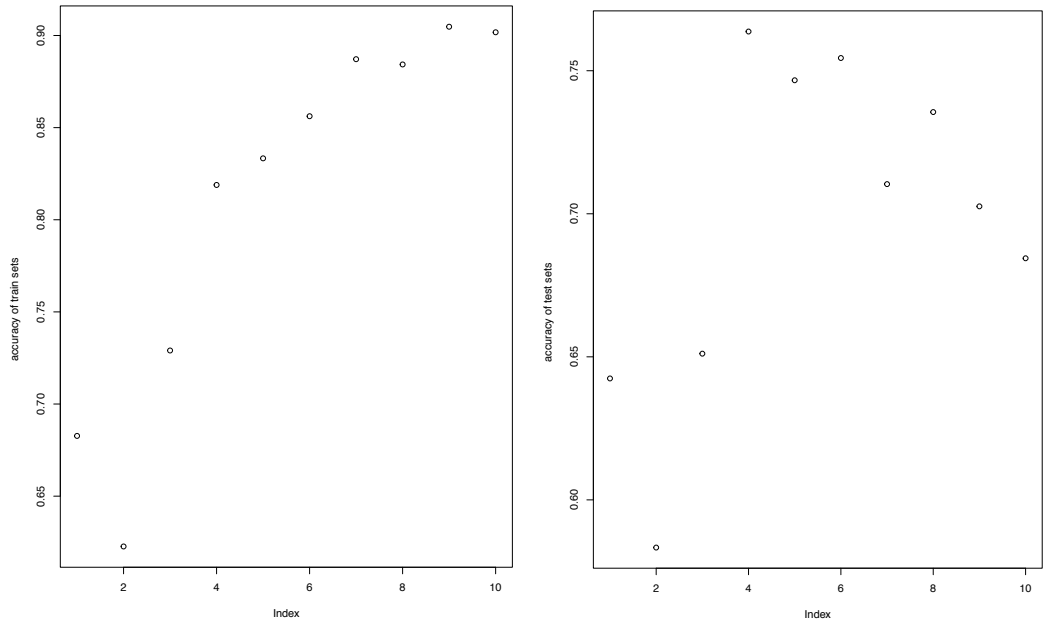


Figure A2) mean accuracy of trained logistic regression (with package nnet) with different number of PCs as input in training set (left) and test set (right)

A1-2-2: MULTINOMIAL LOGISTIC REGRESSION WITH GLMNET PACKAGE.

In another try to fit logistic regression for classifying MPN stages based on the first 10 PCs, we tried another package called *glmnet*[200] which its providers claim it includes extremely efficient procedures for fitting the entire lasso or elastic-net regularization path for linear regression, logistic and multinomial regression models, Poisson regression and the Cox model. Two recent additions are the multiple-response Gaussian and the grouped multinomial regression[200].

A1-2-2A: DATA PREPROCESSING AND FORMATTING

In order to use logistic regression routine in the *glmnet* package, we again first ran PCA on normalized gene expression and then selected the first 10 PCs to train a logistic regression model. Here we don't need any special encoding of labels as we needed for the *neuralnet* function in the neuralnet package. And a vector which codes different stages of MPN for each patient with numbers is enough.

A1-2-2B: APPLYING LOGISTIC REGRESSION (GLMNET) FOR CLASSIFICATION ON PCS

The *glmnet* package is very convenient to use. For fitting the model, we used the function *glmnet.* For predicting the class labels for new data, we used the function *predict.* To install the glmnet package, we need to run the following commands in Rstudio:

```
> install.packages("glmnet")
>library (glmnet)
```

As already mentioned, there is no need for any special format. The input of the *glmnet* function can be in a format of a matrix. We should provide predictors (PCs), response (class labels) and the type of model which we wish to fit, here "*multinomial".*

Again, I split randomly the data in 70% training and 30% test set to make it possible to compare with the previous results. Then we trained 100 models separately for every training set. This routine is repeated for a different set of inputs, which is a different number of PCs, fed into the logistic regression model (starting from 2 PCs to 10 PCs). Each time, the accuracy of prediction of labels in both of the training and the test set is calculated. For each set of input PCs therefore, we have 100 models and their corresponding accuracies, which again is summarized as the arithmetic mean. The results are shown as the accuracy of the test and the training sets in figure A3. The index in both sides of figure A3 indicates the number of PCs used to train the logistic regression model starting from two to 10. Because this package was not accepting only one PC as the predictor, we fed at least two PCs as the input data. Therefore, here we have at most 9 data points. As you can see the model accuracy of training set is quite lower compared with the result of NN, and maximum accuracy is less than maximum accuracy of *nnet* logistic regression models. Unlike NN, the accuracy in the test set is not much different from the accuracy of the training set. Similar to NN and logistic regression model from the *nnet* package, increasing the number of PCs as input increases the classification accuracy in the training set. However, we can see that the highest accuracy (almost 72%) in the test set is achieved, when we use 6 first PCs as the input. Since the test accuracy with 5 PCs is less than with 4 PCs, we concluded that the 4 first PCs is the best set of the input variables to classify the test sets labels.

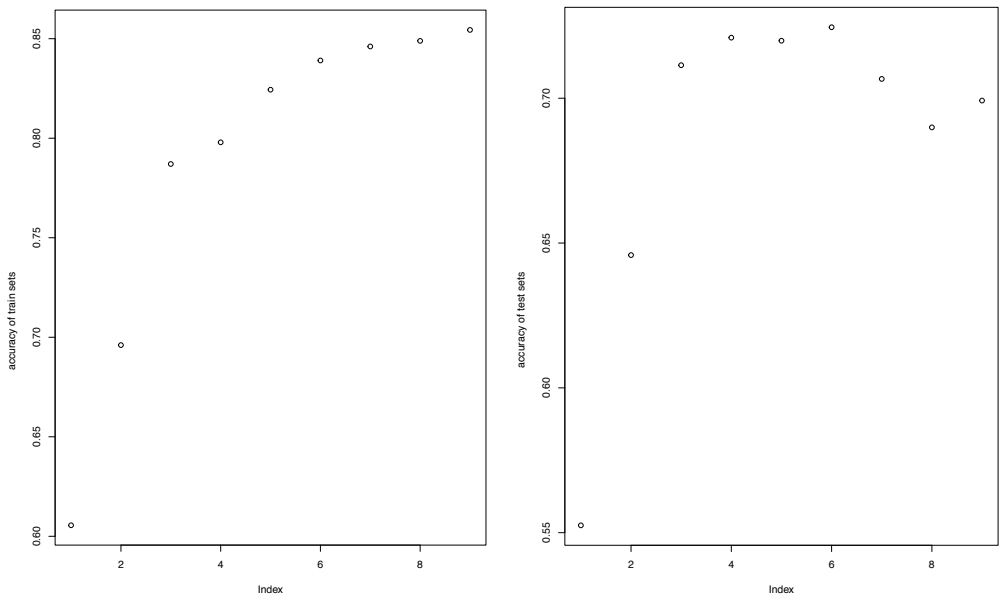


Figure A3) mean accuracy of trained logistic regression (with package glmnet) with different number of PCs as input in training set (left) and test set (right)

A1-3: CLASSIFICATION WITH SVM

In another try to compare different classifiers, we applied the support vector machine. We tried routines from the *liquidSVM*[201] in Rstudio. SVM is innate a binary classifier and when we use it as a multiclass classifier, one-versus-all or all-versus-all reductions to binary classes should

be provided for calculation of the lost function. To be able to use the liquidSVM package we have to first install this library:

```
install.packages("liquidSVM")

library(liquidSVM)
```

In order to use the Multiclass SVM function of package *liquidSVM,* we again firstly ran PCA on normalized gene expression and then selected the first 10 PCs. Here there is no need for any special encoding of labels as we needed for the *neuralnet* function in the *neuralnet* package. Therefore, a vector which encodes the different stages of MPN for each patient with numbers is enough. We used the function *svmMulticlass.* It accepts at least two inputs, first the predictors, here PCs, and second the class labels. There can be a third option which decides if we use the strategy of one versus all comparison or all versus all. We should also set the cost function which can be either hinge or least square.

Again, the whole data is randomly split into 70% training and 30% test to make it possible to compare with the previous results. As for previous models' training procedures, we trained 100 models separately for every training set. We repeated this routine for a different number of inputs, which is a different number of PCs, feeding the SVM model (starting from only with 2 PCs to 10 PCs). Each time the accuracy of prediction of labels in both of the training and the test set is calculated. For each set of input PCs, therefore, there is 100 models and their corresponding accuracies, which again is summarized as the arithmetic mean. The results are shown as the accuracy of the test and the training sets in figure A4.

The index in both sides of figure A4 indicates the number of PCs used to train the logistic regression model starting from tow to 10. Because this package was not accepting only one pc as predictors, we fed at least two PCs. Unlike the previous methods, we don't see a special routine of improvement of the accuracies as the number of PCs increases. The highest accuracy in the training set is achieved when we use all of the 6 PCs and it is about 88%. The highest accuracy in the test set is when we use 7 PCs as predictors for predicting class labels of each patient and it is about 77.5%.

A1-4: Conclusion of Gene expression classification

In this part, we tested different algorithms on a different number of PCs of gene expression profile to select the best algorithm and best number of PCs for the classification of MPNs. We could observe that although NN can fit with very high accuracy (92-98%) on the training set, the predicted labels for the test sets are significantly less accurate (49%-73%). Among the rest of the three algorithms, the logistic regression model from the *glmnet* package and the *SVM* classifier had the same behavior (maximum test accuracy for both is between 70% and 77%). So, the best classifier on gene expression principal components would be either logistic regression model or SVM model. Another aim of this section was to select the right number of PCs. Based on which method is being used to train the classifier, the best number of PCs is various. As the best classifier is either the SVM or the logistic regression classifier, and the best number of PCs is either 6 or 4 respectively. We continued our further analysis with the logistic regression classifier with the first four PCs as input.

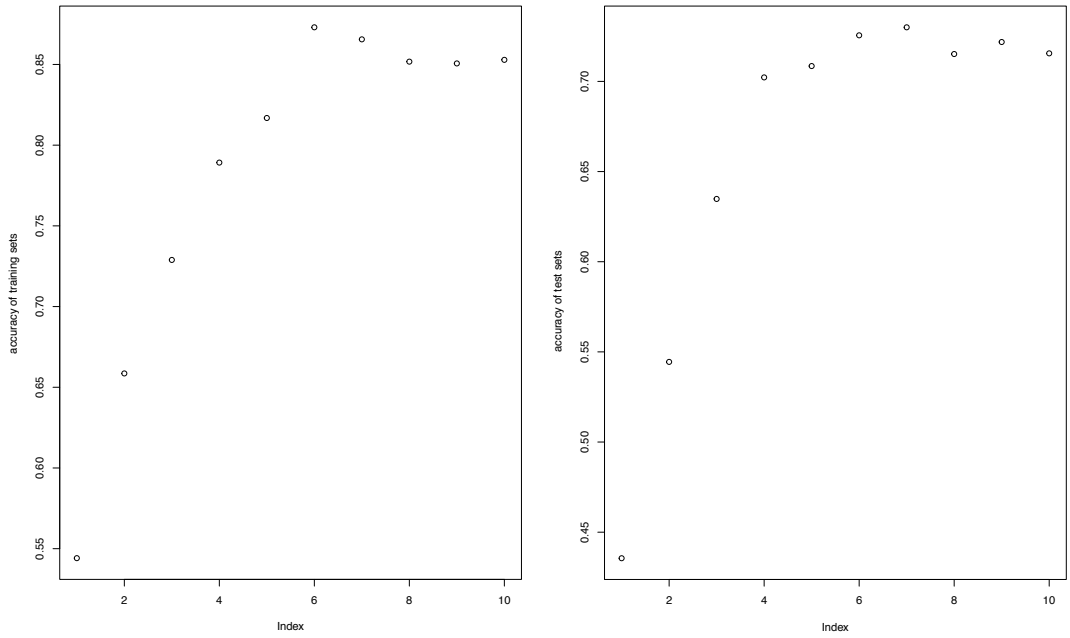


Figure A4) mean accuracy of trained SVM with different number of PCs as input in training set (left) and test set (right)

A1-5: APPLYING PHYSIOSPACE CONCEPT ON GENE EXPRESSION DATA

In section 5-1-4, we have used the concept of PhysioSpace for discovering behavioral changes of different MPNs across physiology-based dimensionalities. PhysioSpace is a method introduced by Lenz and colleagues[167], which relates gene expression experiments from heterogeneous sources using shared physiological processes. Here the focus is to compare the gene expression profile of patients in different stages of MPN with different cell types and tissues to find out which cell type or tissue signature is more present in the progression trend of MPNs. To obtain our goal we used a data set produced by Lukk and colleagues.[202] This dataset is a global map of human gene expression which contains more than 5000 samples in 369 different physiological groups. Samples are taken either from particular cell lines or human tissues. The concept of PhysioSpace is based on the difference or distance of a sample's gene expression from a general reference. PhysioSpace is calculated on the Lukk dataset and consists of 369 axes with a general reference. Each axis has a physiological meaning. As an example, in a gene expression dataset, if we consider each gene as one vector of samples, CD34+ axis in PhysioSpace is a linear combination of most significantly expressed genes in CD34+ tissues, when compared to a general reference. Therefore, if we map a sample on this axis, we actually calculate how significant the CD34+ signature is present in this sample. For clarification, the detailed mathematical details of calculating PhysioScores are described here with an example. In our study in which the focus is on MPNs, we have four different distribution of samples based on their disease stage. Here as an example, consider ET samples distribution to be compared with healthy samples distribution. We call this as physiological process1 (P1). This difference then will be mapped on each axis of PhysioSpace, here in this example CD34+ axis, which we call as physiological process 2 (P2). So, the workflow will be as follow:

P1: physiological state control (state A) => physiological state ET (state B)

P2: physiological state control (state C) => physiological state CD34+ (State D)

We can calculate desired PhysioScores in the following procedure:

Step 1: Calculating the distance /difference between two processes

Calculate the differential expressions for all genes x between states A (x_A) and B (x_B):

$$P_{AB} =: \quad p(A=>B) = -\log10(p(x_A,x_B))*sign(<x_B>_B- <x_A>_A)$$

We do the same for the second process (P2):

$$P_{CD} =: \quad P(C=>D) = -\log10(p(x_C,x_D))*sign(<x_C>_C- <x_D>_D)$$

The p-values $p(x_C,x_D)$ and $p(x_A,x_B)$ can be calculated using t-test.

Step 2: The selection of most significantly expressed genes:

We select the 5% most upregulated genes in P_{AB}, and call them *u*. We select the 5% most downregulated genes in P_{AB} and call them *d*.

Step 3: Mapping of Process P1 on Process p2:

We calculate the physio score of p_{phys} (P1=>P2) which means how similar P1 is to P2 by the following formula:

$$P_{phys} = -\log10(p(P_{CD}(u), P_{CD}(d))) *sign(<P_{CD}(u)>- <P_{CD}(d)>)$$

In step 3, the p-values $p(P_{CD}(u), P_{CD}(d))$ must be calculated using *Wilcoxon-Ranksum test*, as we are interested the how up/downregulated genes are sorted.A *t-test* is not sufficient.

When considering different processes as Process 2, in the end, we have 369 PhysioScores respective to 369 on PhysioSpace axes for each ET patient.

A2: PROCESSING GSG-MPN DATA SET

In this section, we are interested to see if the clinical data excluding gene expression is putative of the classification of the different MPN stages. For our purpose, we used data from GSG-MPN library to apply different types of pattern recognition machinery on. Candidate models are the NN, the SVM, and the logistic regression models. Again, as the preprocessing step for the gene expression analysis, we do cross-validation for class prediction and based on the best result, we select our model. Dividing the data by 70% as the training set and 30% as the test sets for 100 times. We do the cross-validation each time and calculate the classification accuracy. We summarize then the resulted accuracies in the arithmetic mean for the final model selection.

A2-1: DATA PREPROCESSING:

The GSG-MPN dataset has 572 patients, in 5 different categories of MPN: ET, PV, PMF and Post ET/PV myelofibrosis. The registry aimed to collect information in 220 different variables, including complete blood counts, mutation profile (if there are any mutations and the type of existing mutations), treatment profile and occurring complications such as bleeding, thrombosis and so on. However, data is extremely sparse, leaving most of the covariates with a rare number of samples. Therefore, we need to refine and shrink the dataset so that there is enough number of samples for processing techniques. Fortunately for most of the samples,

the complete blood count (CBC) reports are available. CBC reports include information about the total number of leukocytes, the ratio of WBC cell types (such as neutrophil, lymphocytes, monocyte, eosinophil, and basophil), hemoglobin, hematocrit, and platelets. In total, CBC report is available for 450 patients including 150 ET patients, 160 PV patients, 104 PMF patients, 19 post-PV MF patients and 17 post-ET patients. As the first variable in our refined dataset is leukocytes and other WBC cell types are reported as the ratio of this variable, first we need to remove leukocytes and convert the ratios to absolute values in order to make the predictors independent. This independence is required for further processing. It is worth mentioning that the cell counts that we use for further analysis are gathered at the time of entering the registry and not at the time of diagnosis. Many patients are diagnosed long before entering the registry, and many of them are entered in the registry as they are diagnosed as MPN patient. In order to take the most benefit of the data, and because there are only 53 samples out of 450 samples with all refined variables available, prior to all processing methods, we did imputation to increase the number of samples. There are several packages available for imputing missing data. We used *mi*[203] package. The imputation of missing value is done in an approximate Bayesian framework. After imputing and by subtraction, we tested if the new non-missing values are the same as before. The biggest difference was in the range of 10^{-10} which in the context of the original value is considered as zero.

A2-2: TRAINING A CLASSIFIER FOR MPNS BASED ON CELL COUNTS:

A2-2A: TRAINING A NN FOR MPNS CLASSIFICATION BASED ON CELL COUNTS

For training a NN, we should a *priori* set the network architecture, which is the number of hidden layers and the number of nodes in each layer. For this purpose, we did a trial and error process to find a structure which is not only fast enough but also is capable of converging. Starting with a deep NN consisting of three hidden layers with 4 nodes in each layer, the network could not converge with the default maximum number of iterations and we had to increase the number of iterations. By setting the number of iterations to 1e6, although NN converges, it is a very time-consuming process. Then I tried the following structures: three hidden layers with 4,10 and 4 nodes in each layer, three hidden layers with 5,10 and 5 nodes, in each layer. The same problem of not converging in default iteration number occurred and by increasing the number of iterations, the system is slow in converging. Continuing the trial and error process, I came up with two structures which converges fast. One is a shallow network consisting of 10 nodes in its hidden layer and another is a deep NN with two hidden layers of seven and 10 nodes in the first and the second layers respectively. After selecting the NN architecture, I split the data in 70% training and 30% test set and train the network for 100 times. Then I calculated the accuracy of classification for every 100 times. The shallow network was fast but for 100 times run the two-layer neural network needs more time and it takes some hours. The results indicate that regardless of the network structure, the network tends to overfit the data in the training set, therefore the accuracy of the predicted labels in the test sets is very poor. For a shallow neural network with one hidden layer of 10 nodes, the average accuracy is 86.5% in the training set and 58.2% in the test set. For the two-layer NN, the average accuracy is 90.5% in the training set and 56.4% in the test set. The results of the accuracies in the training and the test sets of the two-layer NN are brought in figure A5 as an example.

A2-2B: TRAINING SVM AND LOGISTIC REGRESSION FOR MPNS CLASSIFICATION BASED ON CELL COUNTS

Other candidate algorithms are logistic regression and SVM. The same functions and packages are used here in this part as the one used for the gene expression part. The results

of accuracy in test and training set of the same portioning (30%-70%) in 100 runs are shown in figures A6 and A7. SVM and logistic regression don't significantly differ in terms of the resulted test accuracy. For the training set, SVM performs a bit better but that doesn't lead to a better performance in the test set. Therefore, we picked the logistic regression as the best model. And continue our further processing with logistic regression.

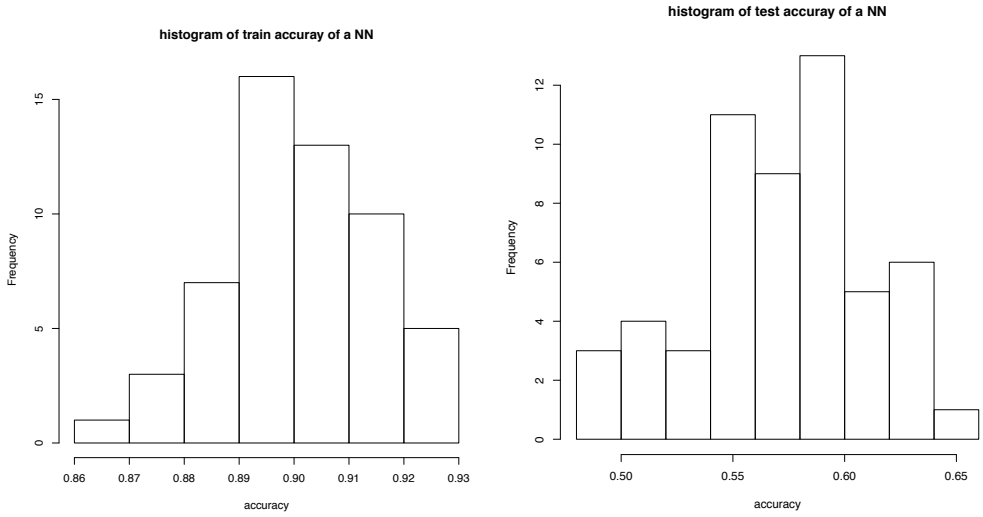


Figure A5) mean accuracy of trained NN with cell count as input in training set (left) and test set (right)

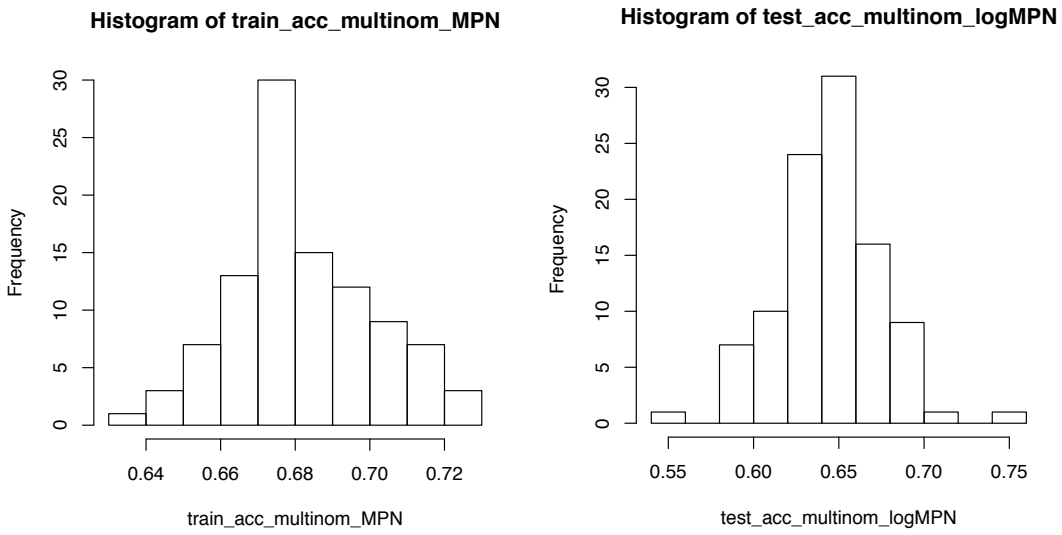


Figure A6) mean accuracy of trained logistic regression classifier with cell count as input in training set (left) and test set (right)

After selecting logistic regression as the best classifier, we trained a logistic regression classifier by considering an ordinal class definition, such that ET samples are grouped as the first/least progressive stage, PVs as the second, PMF as the third, POST-ET MF as the fourth and POST-PV MF as the fifth/most progressive stage of MPN. The results are brought in table A1 as the accuracy of classification. As it is stated in section 5-2-1 and shown in table A1, the accuracy of classifying POST-ET MF and POST-PV MF is zero. Interestingly when we

investigate the predicted labels, we see that out of 19 POST-ET MF samples, 13 are labeled as PMF and majority of the rest are labeled as ET. For POST-PV MF, half of the samples are labeled as PMF and half of them as PV. Therefore, in the next try due to myelofibrotic nature of POST-ET/PV MF, the labels of these groups are changed to 3, meaning that there are three classes of MPN: ET, PV and, MF. Again, a logistic regression classifier is trained on the data with the new labels.

Now with the new labeling, the classification accuracy of MF increases, although there is a decrease in ET prediction accuracy (table A2). I also trained a classifier as there are no ordinal but categorical labels with two modes: original labels and merged labels (POST ET/PV MF are considered MF).

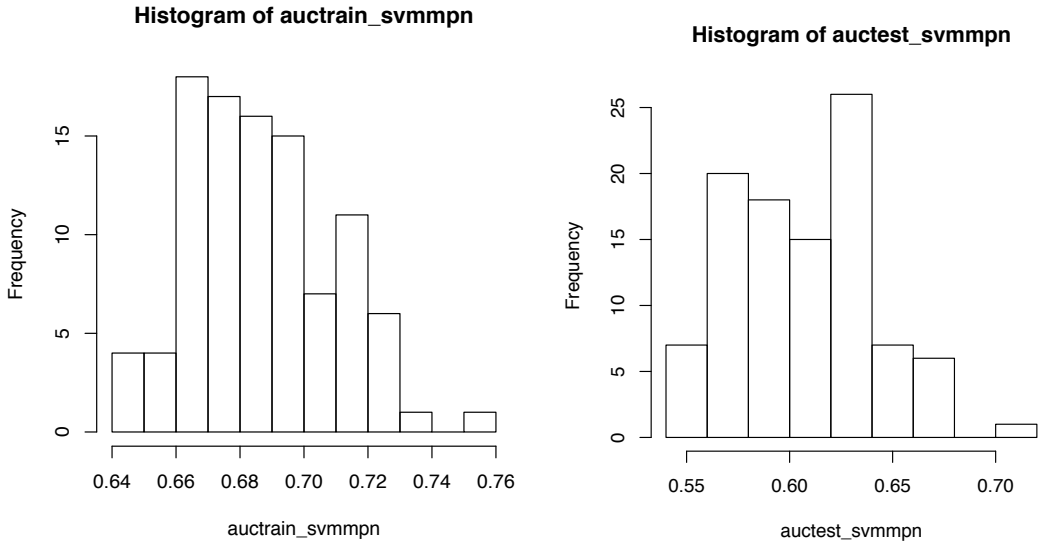


Figure A7) mean accuracy of trained SVM with cell count as input in training set (left) and test set (right)

Table A1: accuracy result of logistic regression classifier on cell counts

MPN type	Classifications accuracy
ET	**80.6%**
PV	**78.75%**
MF	**55.7%**
ETMF	**0**
ETPV	**0**

Table A2: accuracy result of logistic regression classifier on cell counts with merged labels

MPN type	Classifications accuracy
ET	60.33%
PV	78.12%
MF	63.46%
ETMF (as MF)	73.68%
PVMF (as MF)	58.8%

Table A3: accuracy result of logistic regression classifier on cell counts with non-merged categorical labels

MPN type	Classifications accuracy
ET	78%
PV	80.6%
MF	51.9%
ETMF	0%
ETPV	0%

Table A4: accuracy result of logistic regression classifier on cell counts with merged categorical labels

MPN type	Classifications accuracy
ET	76.67%
PV	75%
MF	53.84%
ETMF	78.95%
ETPV	58.82%

Poor accuracy for MF and null accuracy for POST ET/PV MF classifications may imply that cell counts alone don't carry enough information for classification of these stages of MPN diseases. As it can be seen in table A2, changing the ordinal labeling to categorical labeling doesn't improve the classification accuracy of POST ET/PV MF. But merged and categorical labeling improves the overall accuracy in both training set (from 67.8% to 72.25%) and test set (from 64.4% to 69.3%).

A2-2C: TRAINING RANDOM FOREST FOR MPNS CLASSIFICATION BASED ON CELL COUNTS

In addition, we also tested the *Random forest* algorithm on cell count data as another classification method. To apply random forest classifier, we used the *randomForest*[162] package. There is not much better result than logistic regression in the training and the test sets. The average training set accuracy is 68.1% and the average test set accuracy is 68.89%. As it is mentioned in the package description, the *randomforest* function automatically does three-fold cross-validation. However, we part the data by 70% -30% of the training and the test 100 times. and summarized the results as the arithmetic means of accuracies in all 100 times of training.

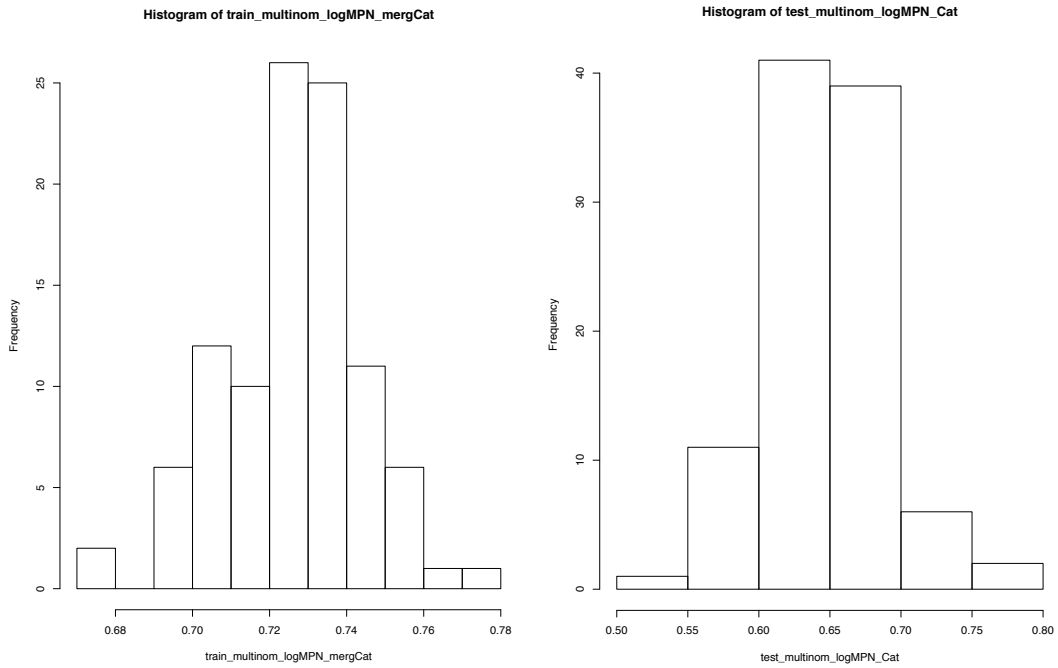


Figure A8) mean accuracy of trained logistic regression classifier with cell count as input and categorical merged labels as output, in training set (left) and test set (right)

A2-2D: RESULT OF MODEL SELECTION FOR CELL COUNT-BASED CLASSIFIER

Training and testing several classification algorithms, we concluded that the logistic regression is the optimal model for classifying MPNs based on cell count data. We also observed that the cell count data is poorly predictive of myelofibrosis and that the secondary myelofibrosis does not deviate from the primary myelofibrosis based on the cell count data. Therefore, we considered all myelofibrosis patients including the primary and the secondary ones in one large category of myelofibrosis for further analysis.

A2-3: TRAINING A REGRESSION MODEL FOR MISCLASSIFICATION SCORE OF CELL-COUNT BASED CLASSIFIER

Here in this section we aim to train a regression model based on non-cell count data for predicting the error of cell-count based classifier in our hybrid model.

A2-3A: PREPROCESSING OF OTHER VARIABLES FOR PREDICTION OF MISCLASSIFICATION SCORE

In the GSG-MPN library, there are other reported variables, such as therapy information, age, lactate dehydrogenase (LDH), mean corpuscular volume (MCV) and etc. But as already mentioned at the beginning of this section, due to the sparsity of the GSG-MPN library, we could only use those variables with a reasonable number of samples. In order to select the right variables most efficiently, we created the heatmap of the GSG-MPN data (figure A10). Red pixels show the unavailability of the data for a clinical variable, whereas white pixels show the availability of the data. The list of the selected variables is shown in table A5. Another preprocessing step is creating dummy variables out of multi-categorial data such as therapy. For example, there are overall 7 different therapy options, each of them coded with a number. Converting therapy codes to a dummy matrix increase the number of variables. Other multi-categorial variables are the cellularity of bone marrow and ECOGC health status. Overall, we have 41 variables including binary and continuous values. Now that we refined the variables, in this section we want to predict the misclassification error (achieved by logistic regression on cell counts) of each sample using the listed variables in table A5.

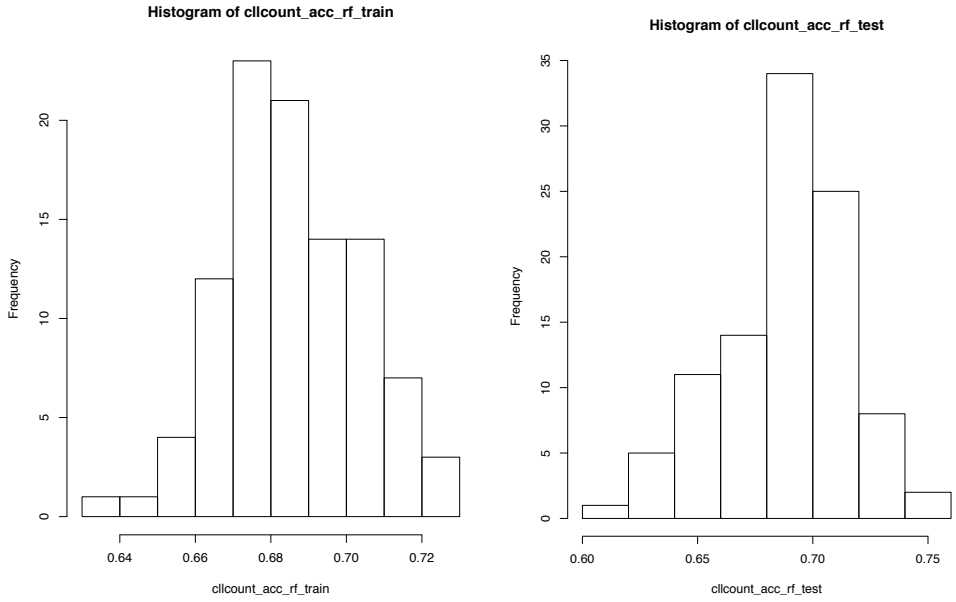


Figure A9) mean accuracy of trained random forest classifier with cell count as input and categorical merged labels as output, in training set (left) and test set (right)

A2-3B TRAINING A NN FOR MISCLASSIFICATION SCORE ESTIMATION

Here we aim to estimate/predict the misclassification error. We use different structures of NN. The quality control of our model is defined as:

$$error = \frac{y - y\prime}{0.5(y + y')} \qquad \textbf{Eq. (12)}$$

In which *y* is the classification error from the cell count-based logistic regression classifier and *y'* is the predicted misclassification error from the trained NN. We considered different architectures of NN such as a three hidden layers network of 10 nodes, a shallow network of 5 to 10 nodes. Due to the big number of inputs, we faced overfitting in training set such that for mean squared error in the training set is almost half of the mean squared error in the test set.

In the other hand due to the nature of input variables (binary), their non-linear dependence and high dimensionality, NN is the first option comes to mind to predict the misclassification error. To overcome the overfitting problem, we applied regularization. The previous neural network packages didn't provide the regularization option, so we use another package called *ANN2*[204] ,which can be installed and utilized through the following codes.

```
> install.packages("ANN2")
>library (ANN2)
```

The function that we use in this package is the *neuralnetwork* fucntion, with the input arguments including input variables, output of the model (here misclassification error), NN architecture as a vector, the type of the modeling problem (regression or classification), the loss function for optimization and type of regularization, which is either ridge or Lasso:

- As the desired output is continuous, we should use a regression model
- We define three hidden layers with 5,10,5 nodes in each layer
- We define the loss function as squared error
- And we use ridge regression as the regulator

So, the function would be like:

```
NN_model<-neuralnetwork(x,y,c(5,10,5), regression=TRUE, loss.type="squared", L=0.5)
```

Table A5: most frequent variables in GSG-MPN registry

variable	type
age	continuous
MCV (mean corpuscular volume)	continuous
CRP (C-Reactive Protein)	continuous
Reticulocytes	continuous
Uric acid	continuous
Bilirubin	continuous
Ferritin	continuous
Therapy1	binary
Therapy2	binary
Therapy3	binary
Therapy4	binary
Therapy5	binary

Therapy6	binary
Therapy7	binary
sex	binary
Cellularity1	binary
Cellularity2	binary
Cellularity3	binary
Cellularity4	binary
ECOGC1	binary
ECOGC2	binary
ECOGC3	binary
ECOGC4	binary
ECOGC5	binary
Statins?	binary
Other anticoagulation	binary
Phlebotomy	binary
Radiation Spleen	binary
Splenectomy	binary
Sonography	binary
Leukocytosis	binary
Acetylsalicylic acid	binary
Thrombocytopenia	binary
Anemia	binary
Malignancy	binary
Severe bleeding	binary
Stem cell transplantation	binary
Other chemotherapy	binary
LDH	continuous
Thrombosis/ thromboembolism	binary
splenomegalia	binary

Here again, we split the data by 70% - 30% for training and test and do the cross-validation for 100 times. Each time we calculated the error from equation 12. We trained our model in three versions depending on the variables feeding the inputs of the network:

- Based on only continuous variables (age, LDH, …)
- Based on only binary variables which their value defines the significantly different distribution of misclassification error
- Based on all variables listed in table A5

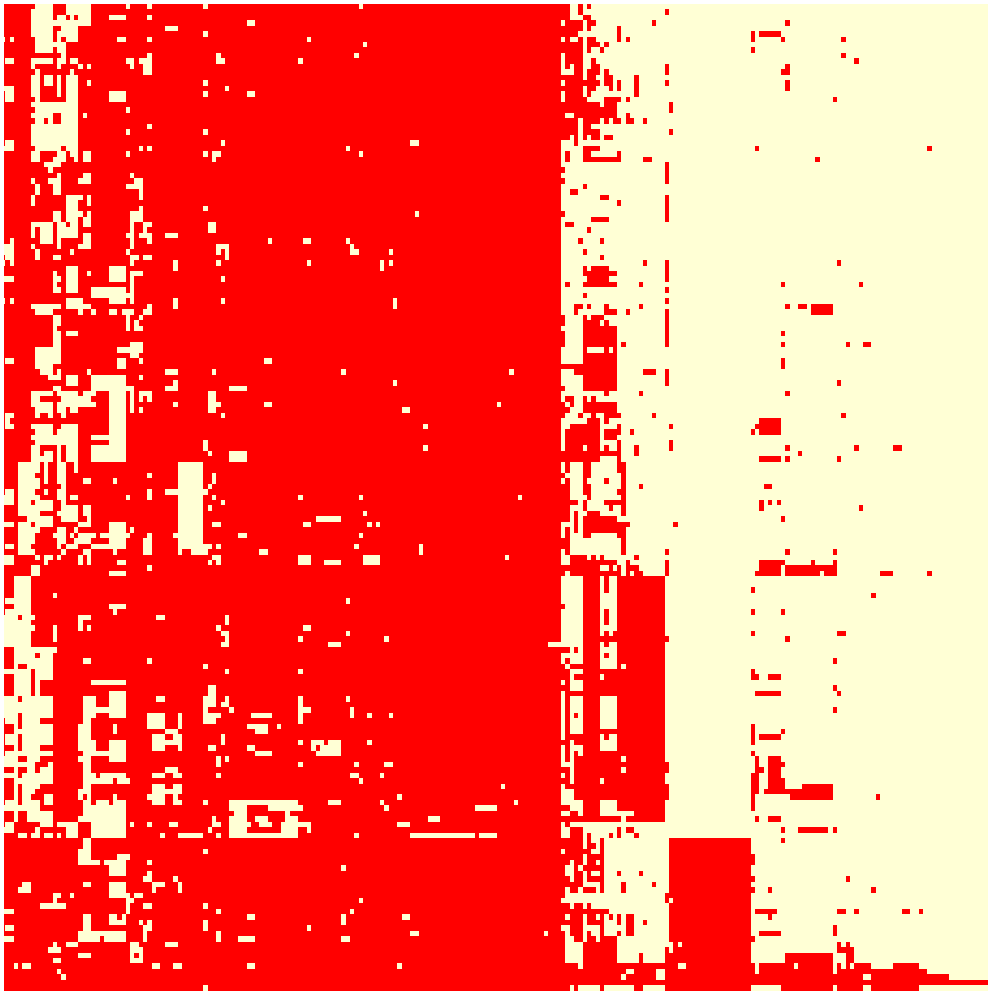


Figure A10) a heatmap of availability of samples (y axis) in each variable (x axis). The red pixel shows no availability and white pixel shows availability of a sample for each variable.

The histogram of errors in the test and training sest are shown in figures A11-A13. Network output doesn't change significantly between different versions of network input. However, the best result is achieved by using all of the variables. Therefore, we used all of the variables as inputs of our network. The mean squared error (MSE) in the training and the test set for each version is listed in table A6.

Table A6: mean squared error of NN prediction with different input settings

Type of inputs	MSE in train	MSE in test
Significant variables	0.488	0.49
Continuous variables	0.502	0.491
All variables	0.486	0.484.

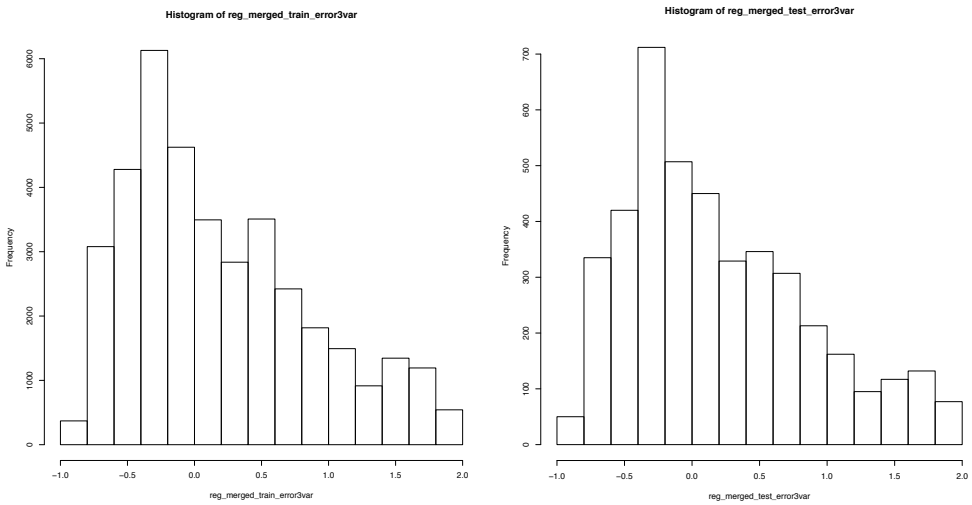


Figure A11) histogram of errors of NN for misclassification score with continuous variables as input in training (left) and test sets (right)

A2-3C TRAINING A RANDOM FOREST FOR MISCLASSIFICATION SCORE

We also checked the random forest as the second black box model for prediction of misclassification error of the first classifier. We used again *RandomForest* Package. Interestingly as most of the variables here are in the binary form, random forest outperforms neural network by less MSE of 0.35 in the training set and 0.2 in the test set (figure A14).

With these results, that the lowest MSE of error prediction is achieved by applying random forest on non-cell count data, we used Random Forest as the second black box in our hybrid model.

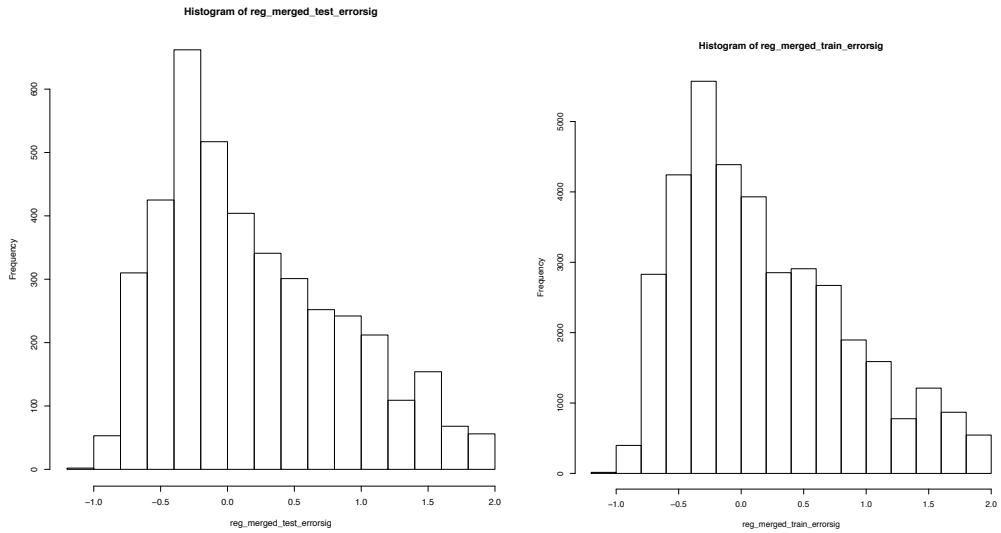


Figure A12) histogram of errors of NN for misclassification score with significant variables as input in training (left) and test sets (right)

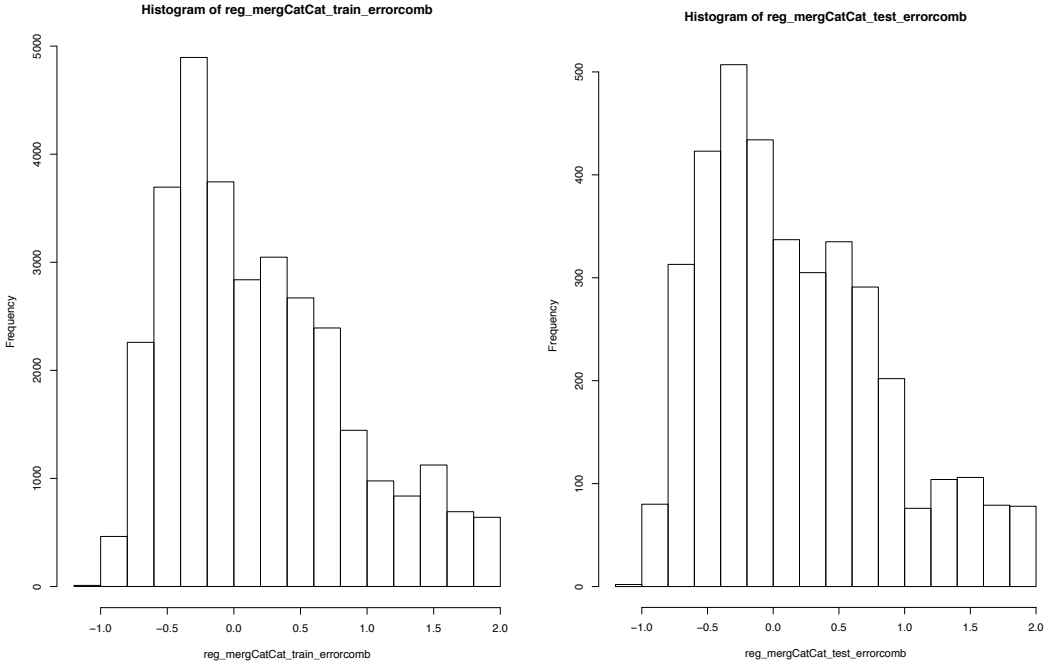


Figure A13) histogram of errors of NN for misclassification score with all variables as input in training (left) and test sets (right)

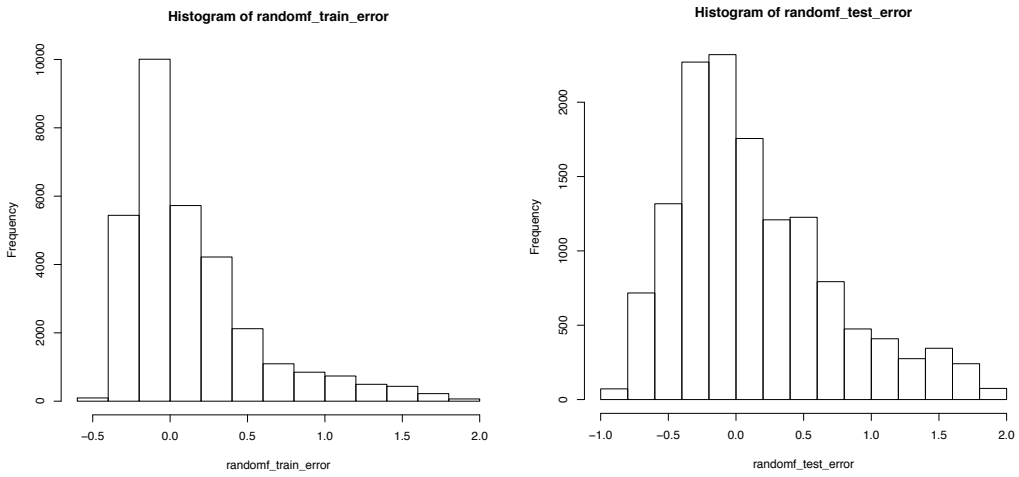


Figure A14) histogram of errors of random forest for misclassification score with all variables as input in training (left) and test sets (right)

A2-4: SELECTING THE BEST MODEL FOR THIRD BLACK BOX

In this section we want to select an algorithm that performs best on the outputs of the two previous black boxes in our hybrid model: first, the cell count-based logistic regression classifier and second the random forest regression model for prediction of misclassification errors of the first classifier. The output of the third black box is again the classes of each sample. Here the goal is to choose the best model among logistic regression, neural network and Random forest as the best classifier. Again, we do the splitting of the data in 70% and 30 % training and test and run the algorithm for the 100 times. The data preparation is the same as before. For training the NN, we did regularization in order to avoid the overfitting. The results are shown as the histogram of accuracy in the training and the test sets, for each algorithm (figures A15-A16).

The overall accuracy of random forest outperforms logistic regression and NN in both of the training and test sets. The overall mean accuracy in the training and the test sets are listed in table A7 for three algorithms.

Table A7: comparison of different models for MPN classification

algorithm	Overall accuracy train	Overall accuracy test
Logistic regression	73.4%	72.2%
Neural network	48.07%	47.4%
Random Forest	82.3%	83.11%

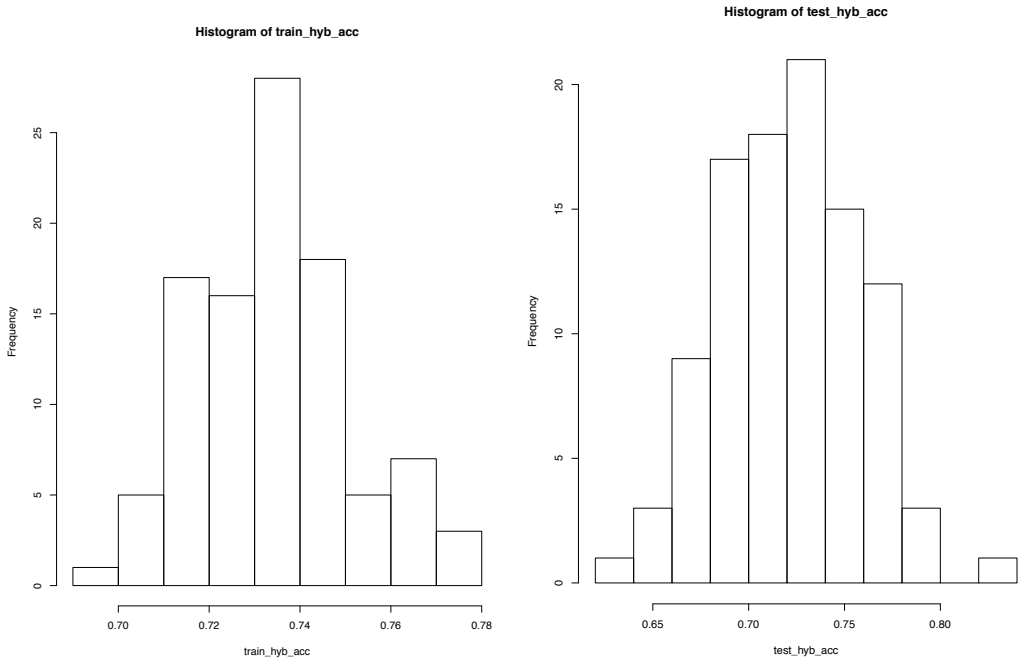


Figure A14) mean accuracy of trained logistic regression classifier with outputs of previous black boxes as input, in training set (left) and test set (right)

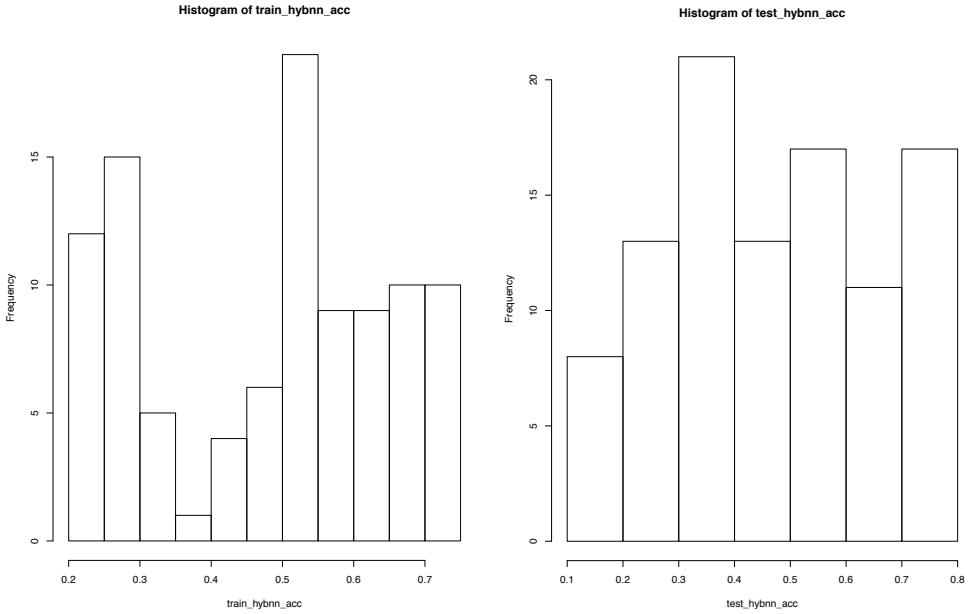
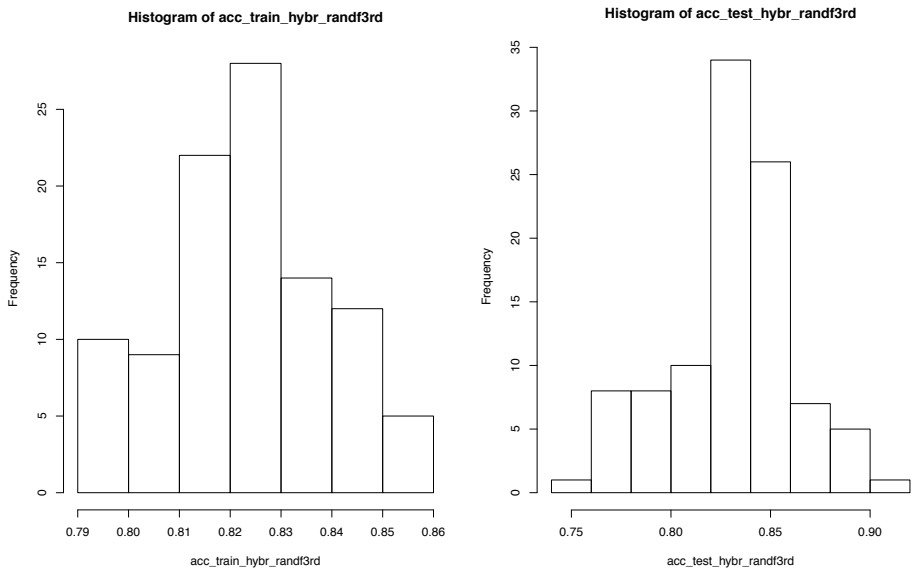


Figure A15) mean accuracy of trained NN classifier with outputs of previous black boxes as input, in training set (left) and test set (right)



FigureA16) mean accuracy of trained random Forest classifier with outputs of previous black boxes as input, in training set (left) and test set (right)

A3: CELL MIXTURE ANALYSIS

Gene expression data are generated from the heterogeneous samples consisting of multiple cell or tissue types, in different proportions[188]. Heterogeneity in sample composition is known as a major confounder in gene expression analysis such as differential expression analysis[205]. As early mentioned in the introduction, in MPNs the number of cell types is changing as a result of mutations or deficiencies in hematopoiesis. Hence, we hypothesized that deconvolution of gene expression of MPN can give insight to the proportion of cells in whole blood samples. For this purpose, we used a package called *CellMix*[188] *developed* by Gaujoux and colleague. The toolbox is available in Bioconductor. This package uses different methods of gene expression deconvolution. It also consists of benchmark dataset form published studies on cell/tissue-specific gene expression or deconvolution method. To deconvolve gene expression, it is important and critical to have access on cell signatures and marker gene sets. *CellMIix* package includes 8 marker gene lists which are compiled from the previous databases[188]. Based on the previous studies [110,140] that suggest a critical role of inflammation and immune system in MPNs and because GSE26049 is a gene expression dataset from the whole blood tissue, we used a marker gene list developed by Abbas and colleagues[206]. In their research, Abbas et al compiled a compendium of microarray expression data for virtually all human genes from six key immune cell types in their activated and differentiated states. They developed Immune Response In Silico (IRIS) as a collection of genes that have been selected for specific expression in immune cells. Fortunately, this gene signature is available in *CellMix* package.

Prior to using *Cellmix* package as it is provided in Bioconductor[207], we need to also install *BiocInstaller*[208] to enable Bioconductor repositories.

```
# install (NB: this might ask you to update some of your packages)
>biocLite('CellMix', siteRepos = 'http://web.cbio.uct.ac.za/~renaud/CRAN', type='both')
>library(BiocInstaller)
>biocLite("GEOquery")
>library(CellMix)
```

The functions to extract cell ratios is *gedBlood* and *coef.*

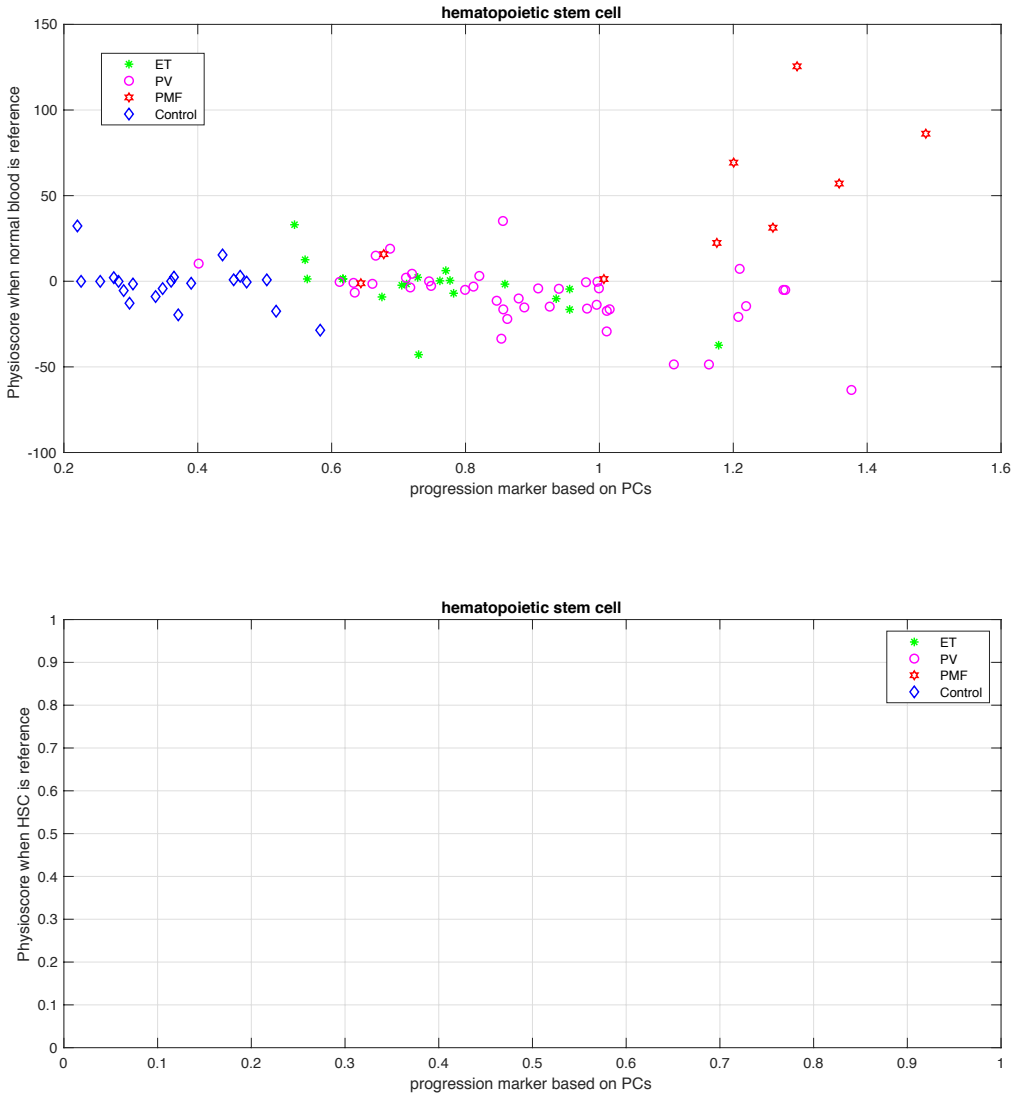
```
gedBlood(expression_gset,verbose=TRUE)
res<-gedBlood(expression_gset,verbose=TRUE)
```

The output *res* is a matrix with 17 rows, with each row representing the ratio of one cell type and, 91 columns with each column representing one sample. As mentioned in chapter 5 (section 5-1-5, figure 41), the number of inputs in the second classifier is 7 and not 17 and that is because we summed over all subtypes of each cell type to reduce the dimensionality and also to make it comparable to PhysioSpace vectors. So, T cell ratio here is the sum of the first four rows of *res.* The first four rows represent T helper cells activated and resting, memory T cells resting and activated. For B cells we summed over the next five rows which represent naïve, active, memory IgG, memory IgA, and memory IgM cells. The 10^{th} row represents the ratio of Plasma cells. 11^{th} row as resting and 12^{th} row as activated Natural Killer (NK) cells are summed together to make the whole ratio of NK cells. Resting and activated monocytes ratios are stored in 13^{th} and 14^{th} rows of *res* matrix and therefore summed together to make the whole ratio of monocytes. The resting and the activated dendritic cells are stored in 15^{th} and 16^{th} cell rows and added together to make the whole ratio of dendritic cells. The last row shows the ratio of neutrophils.

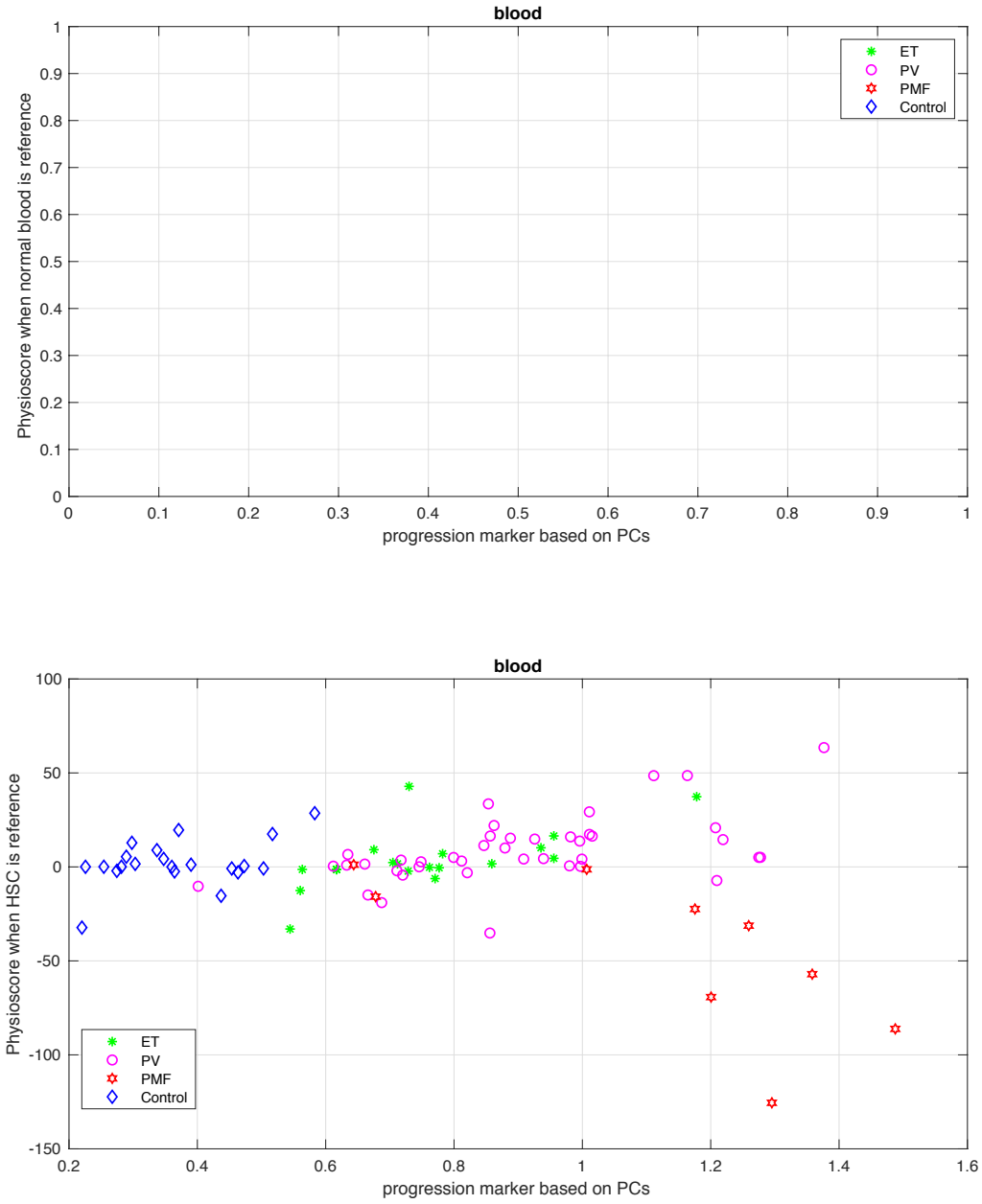
SUPPLEMENTARY FIGURES

The following figures are the results of PhysioSpace analysis on GSE26049 datasets.

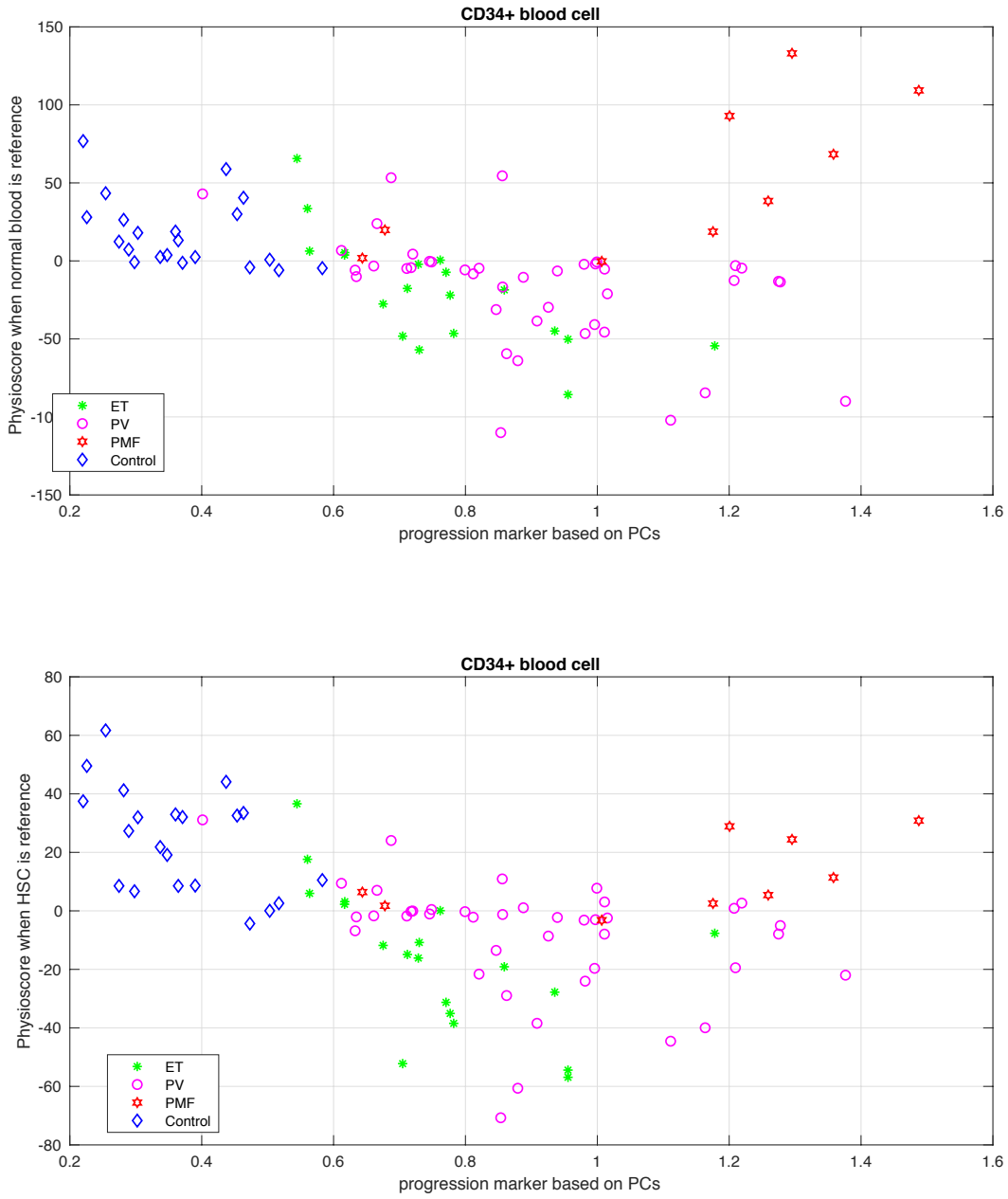
In all figures, the x-axis shows the PC-based progression marker value for each patient, and the y-axis is the calculated PhysioScores on the corresponding physio axes of the coordinate systems with either blood as reference (upper panel) or HSC as a reference (lower panel).



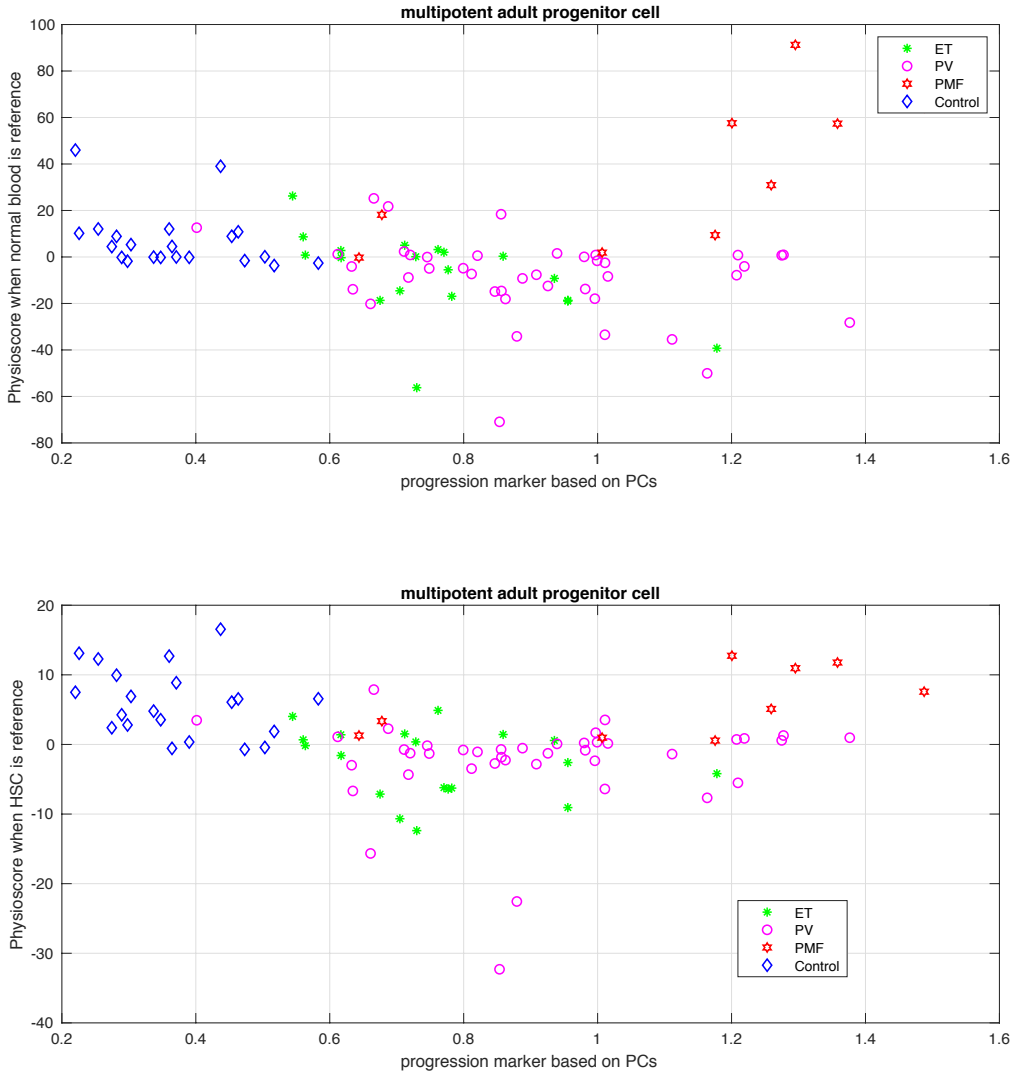
Sfig1: patients' PhysioScores on HSC physio axis in a coordinate system with either blood as the reference (upper figure) or HSC as the reference (lower figure). In the lower figure since the HSC itself is both the reference and the axis, the PhysioScores are zero.



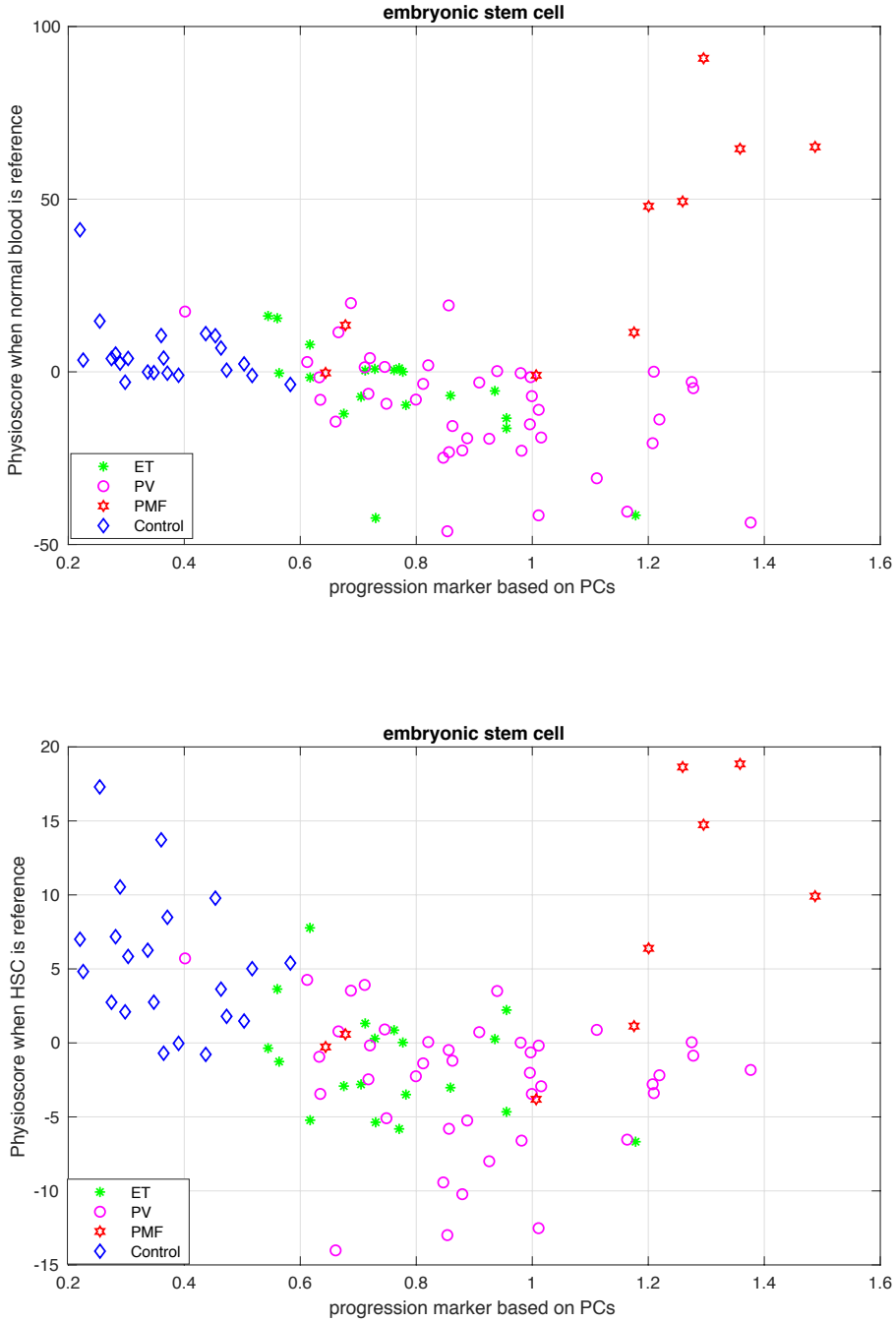
Sfig2: patients' PhysioScores on Blood physio axis in a coordinate system with either blood as the reference (upper figure) or HSC as the reference (lower figure). In the upper figure since the blood tissue itself is both the reference and the axis, the PhysioScores are zero.



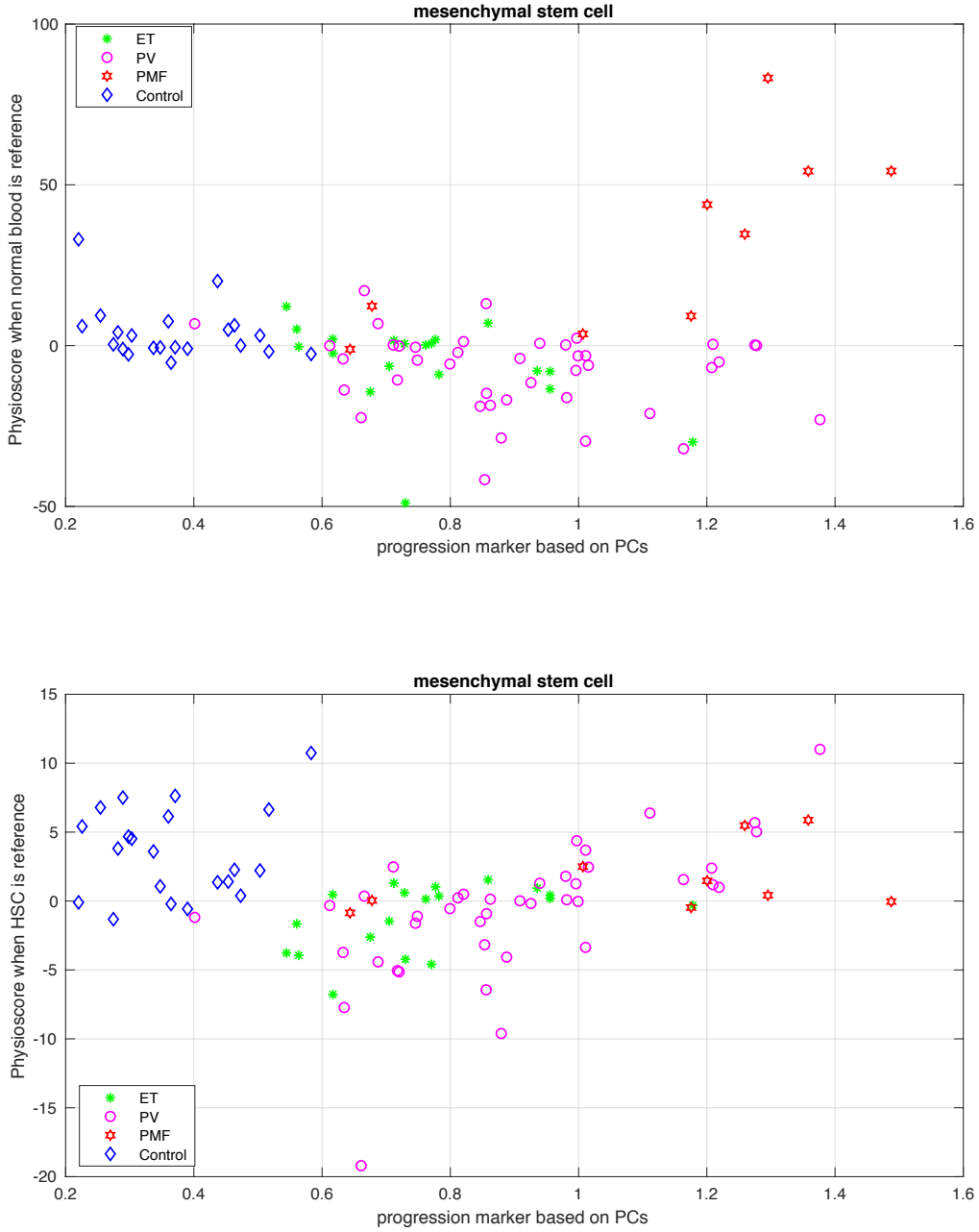
Sfig3: patients' PhysioScores on CD34+ blood cell physio axis in a coordinate system with either blood as the reference (upper figure) or HSC as the reference (lower figure). The positive scores of PMF indicate their stemness. Also, the monotonic decrease of ET samples is observable.



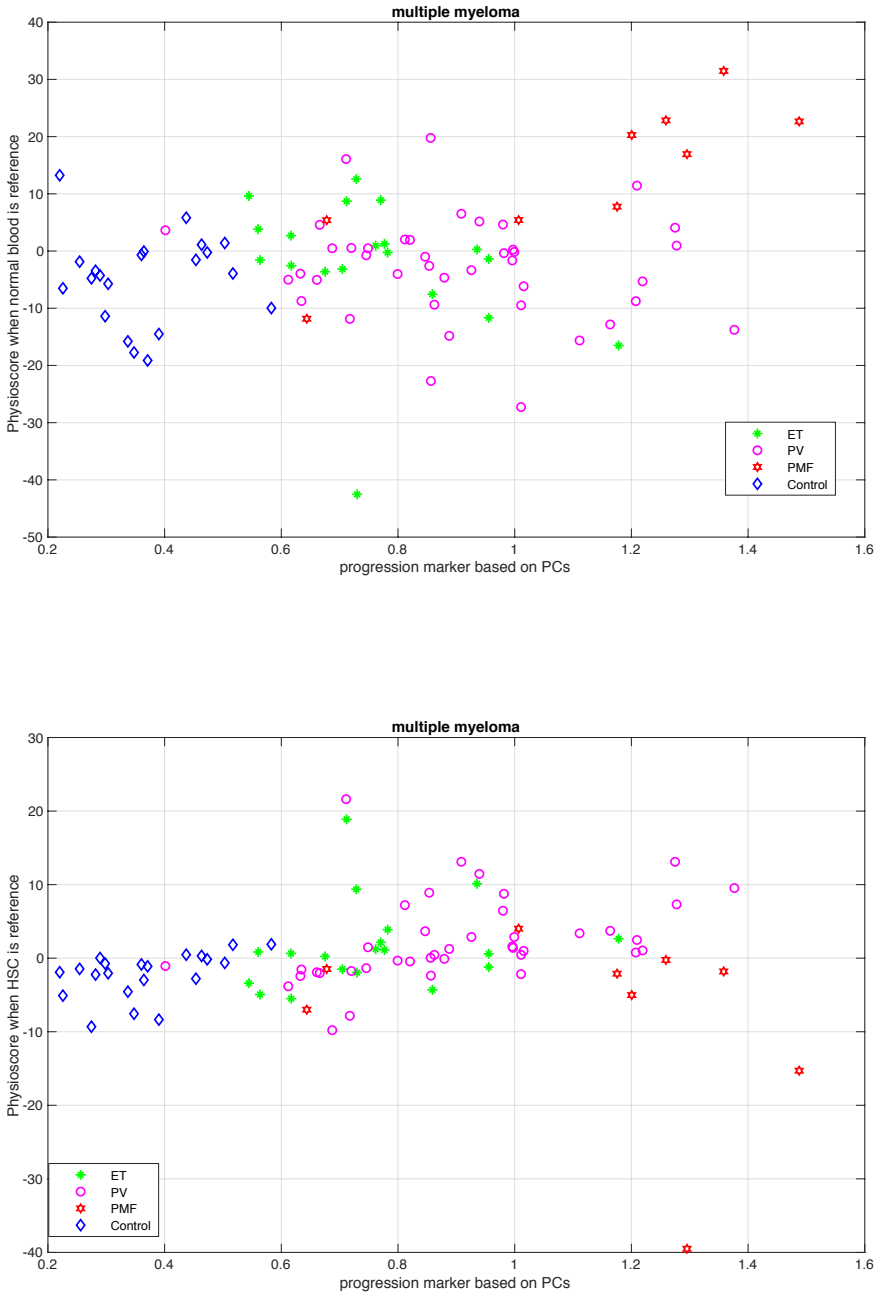
Sfig4: patients' PhysioScores on multiple adult progenitor cell physio axis in a coordinate system with either blood as the reference (upper figure) or HSC as the reference (lower figure). The positive scores of PMF indicate their stemness.



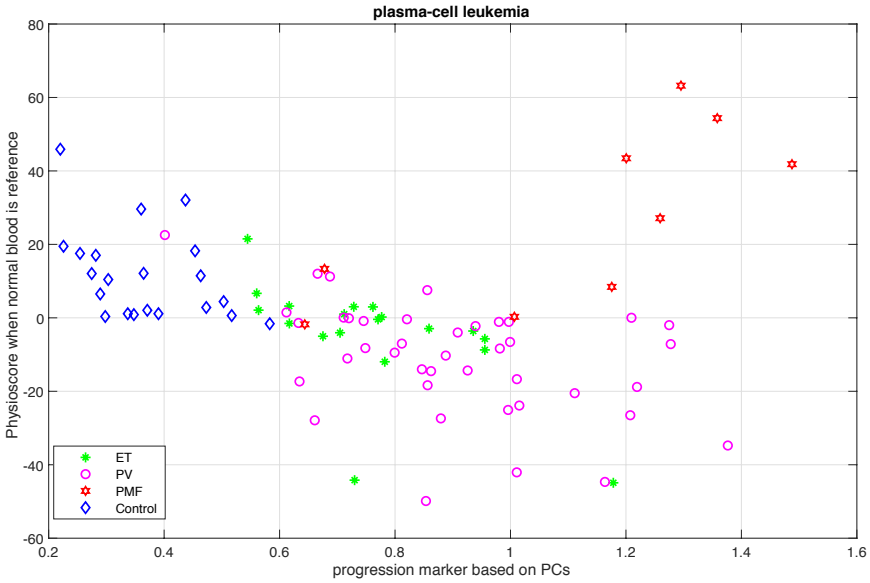
Sfig5: patients' PhysioScores on embryonic stem cell physio axis in a coordinate system with either blood as the reference (upper figure) or HSC as the reference (lower figure). The positive scores of PMF indicate their stemness. Also, the monotonic decrease of ET samples is observable.

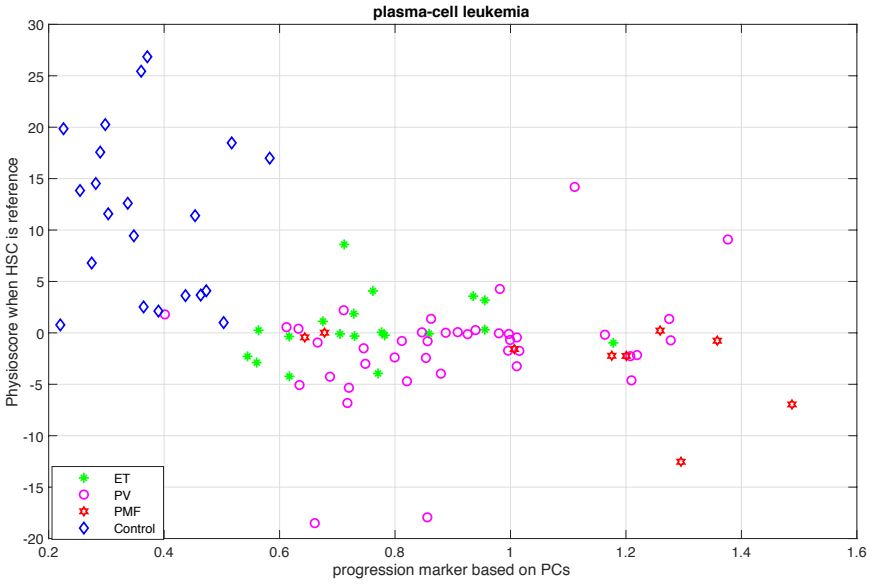


Sfig6: patients' PhysioScores on mesenchymal stem cell physio axis in a coordinate system with either blood as the reference (upper figure) or HSC as the reference (lower figure). The positive scores of PMF indicate their stemness.

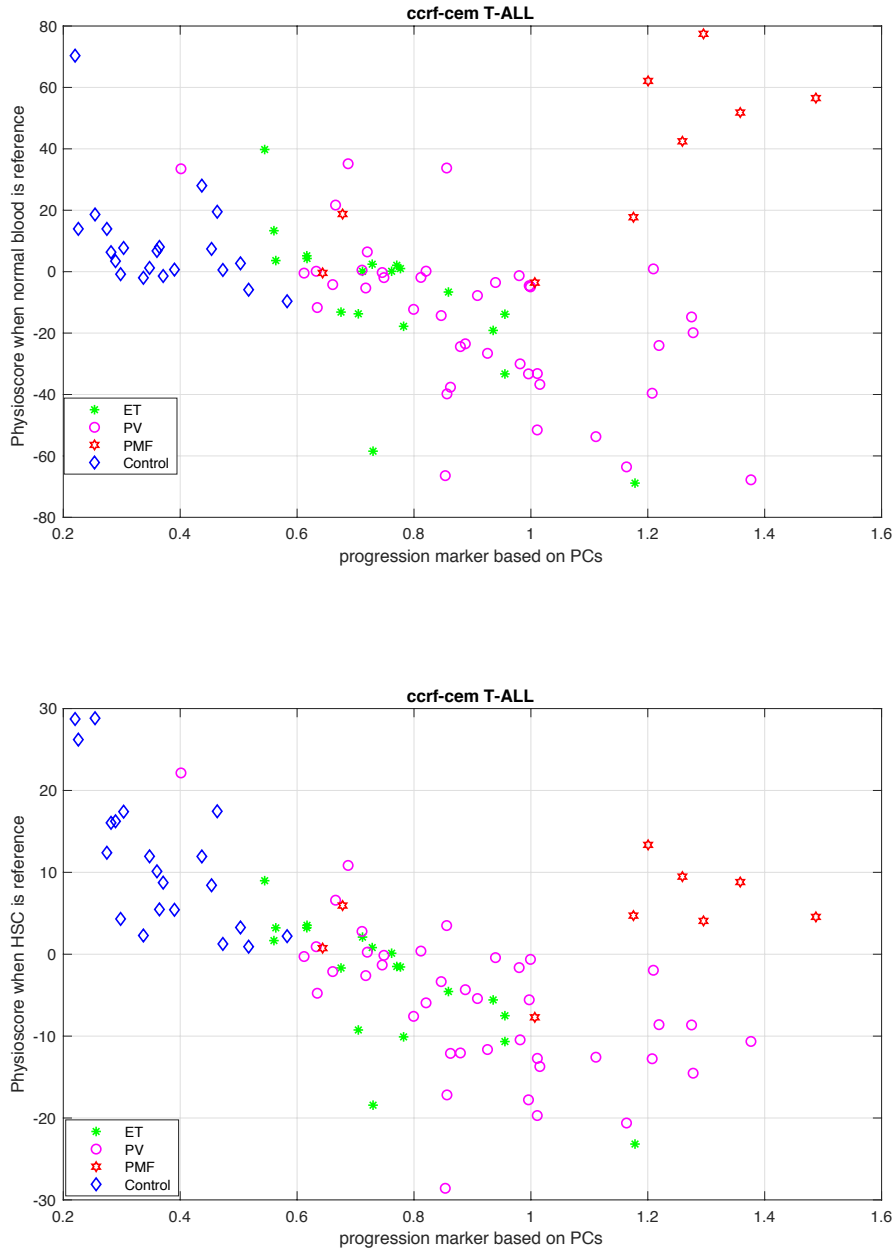


Sfig7: patients' PhysioScores on multiple myeloma physio axis in a coordinate system with either blood as the reference (upper figure) or HSC as the reference (lower figure). The positive scores of PMF on the upper panel show that when blood is the reference of the coordinate system, as we are moving in the direction of multiple myeloma, PMF patients gain higher similarity scores. When HSC is the reference of the coordinate system, as we are moving in the direction of multiple myeloma, PMF patients gain negative similarity scores which means they move against multiple myeloma and preserve their stemness and similarity to HSCs.

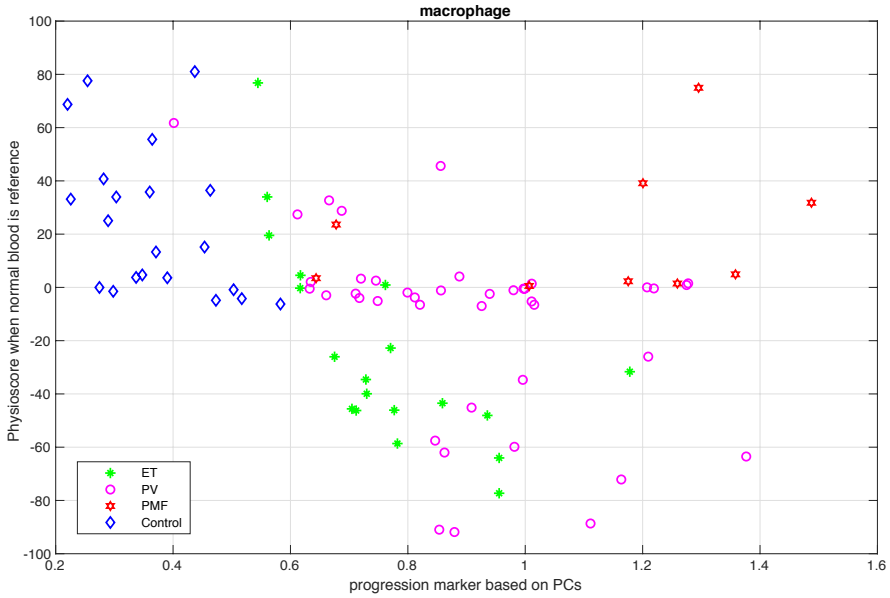


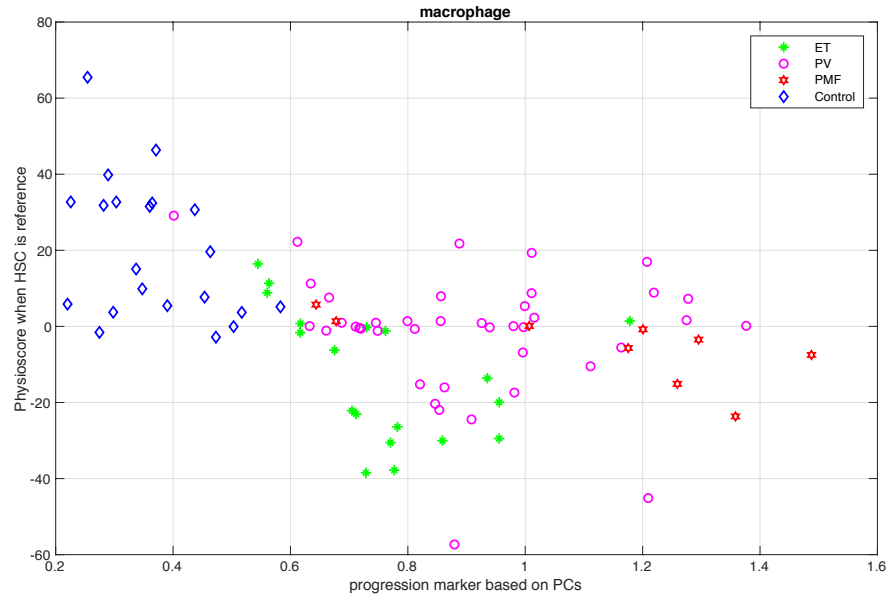


Sfig8: patients' PhysioScores on plasma cell leukemia physio axis in a coordinate system with either blood as the reference (upper figure) or HSC as the reference (lower figure). The positive scores of PMF on the upper panel show that when blood is the reference of the coordinate system, as we are moving in the direction of Plasma cell leukemia, PMF patients gain higher similarity scores. When HSC is the reference of the coordinate system, as we are moving in direction of Plasma cell leukemia, PMF patients gain negative similarity scores which means they move against Plasma cell leukemia and preserve their stemness.

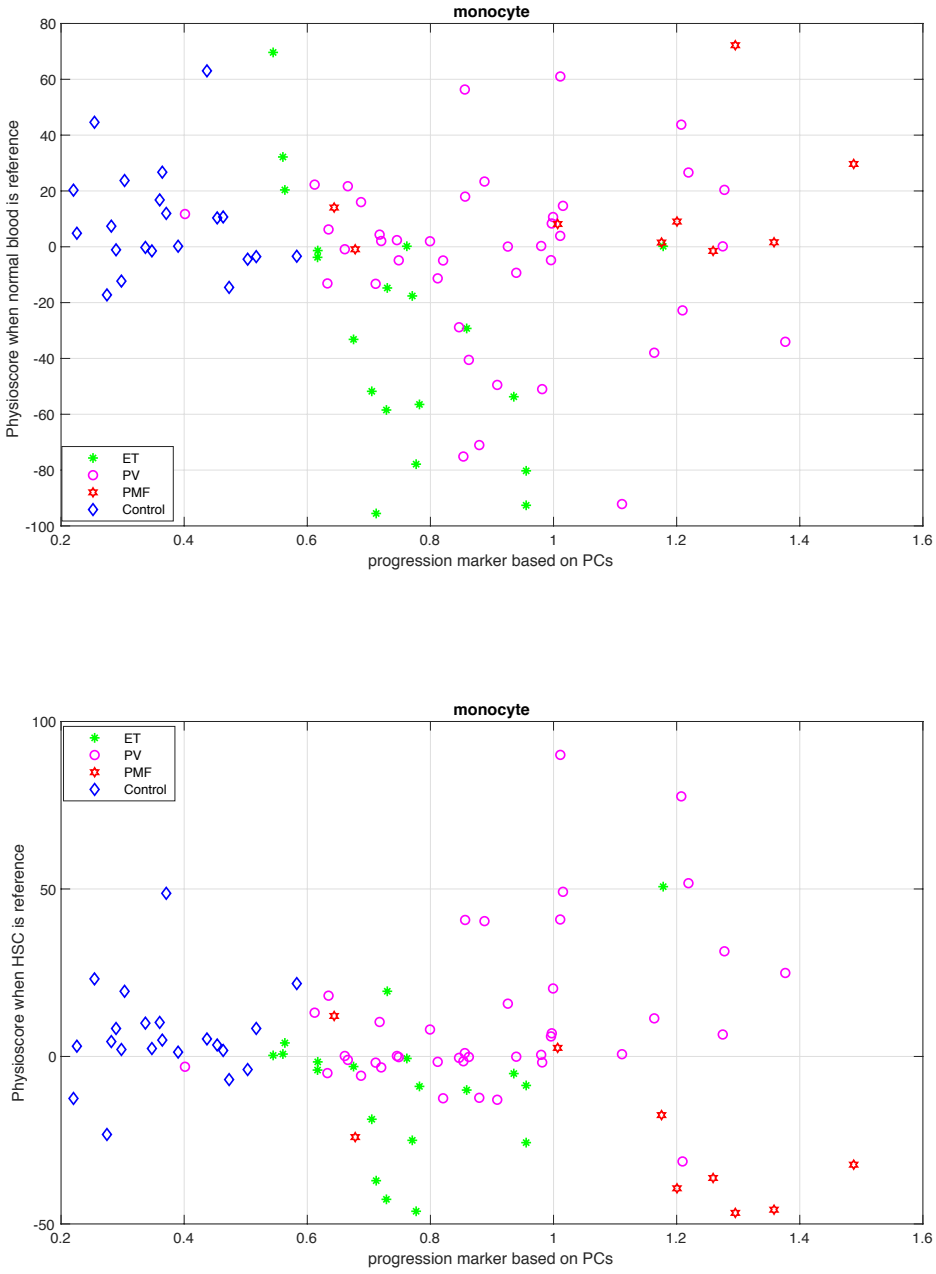


Sfig9: patients' PhysioScores on T cell Acute Lymphoid Leukemia (TALL) physio axis in a coordinate system with either blood as the reference (upper figure) or HSC as the reference (lower figure). The positive scores of PMF patients on the both upper and lower panels show that no matter what the reference of the coordinate system is, as we are moving in the direction toward TALL, PMF patients gain higher similarity scores. When HSC is the reference of the coordinate system, as we are moving in the direction of TALL, PMF patients gain negative similarity scores which means they move against TALL and preserve their stemness. Also, the monotonic decrease of ET samples in both panels is observable.

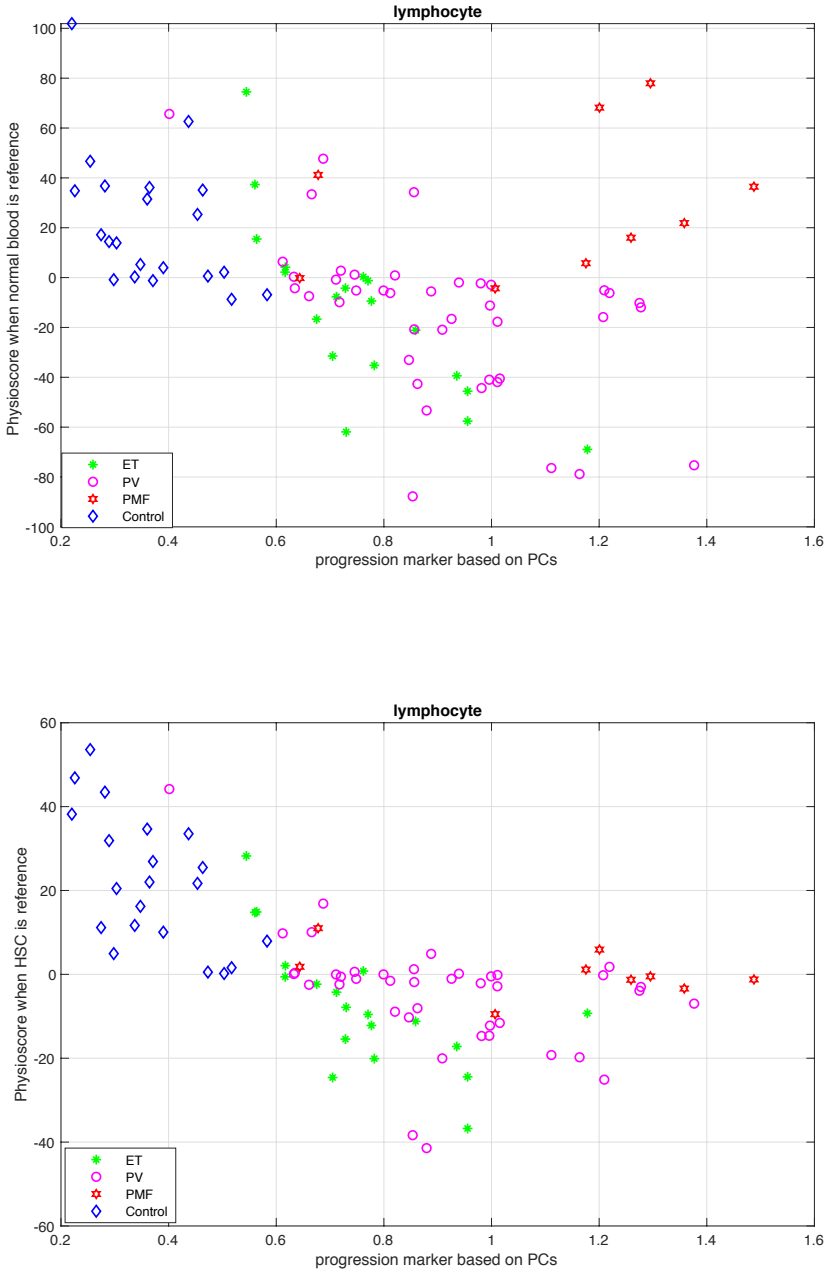




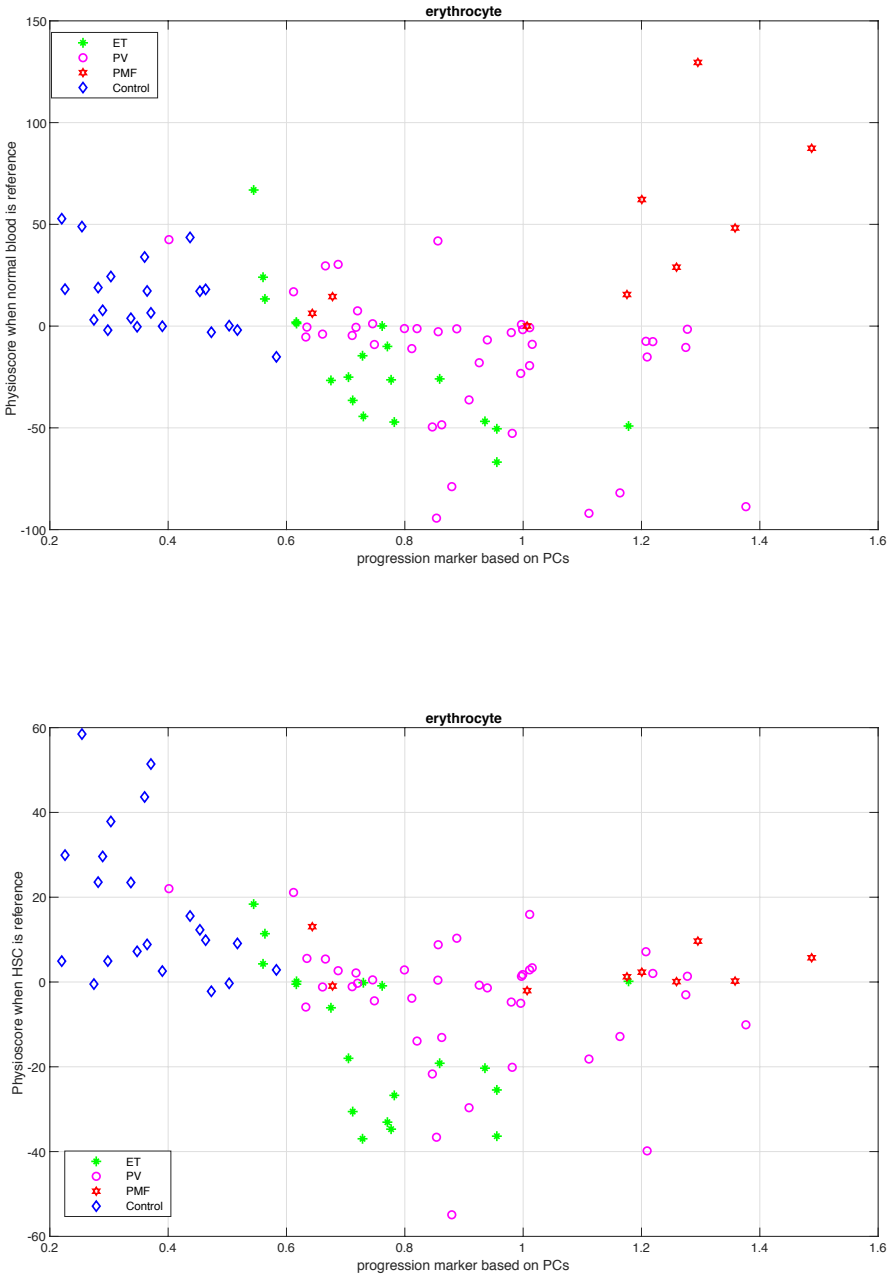
Sfig10: patients' PhysioScores on Macrophage physio axis in a coordinate system with either blood as the reference (upper figure) or HSC as the reference (lower figure). The monotonic decrease of ET samples is observable.



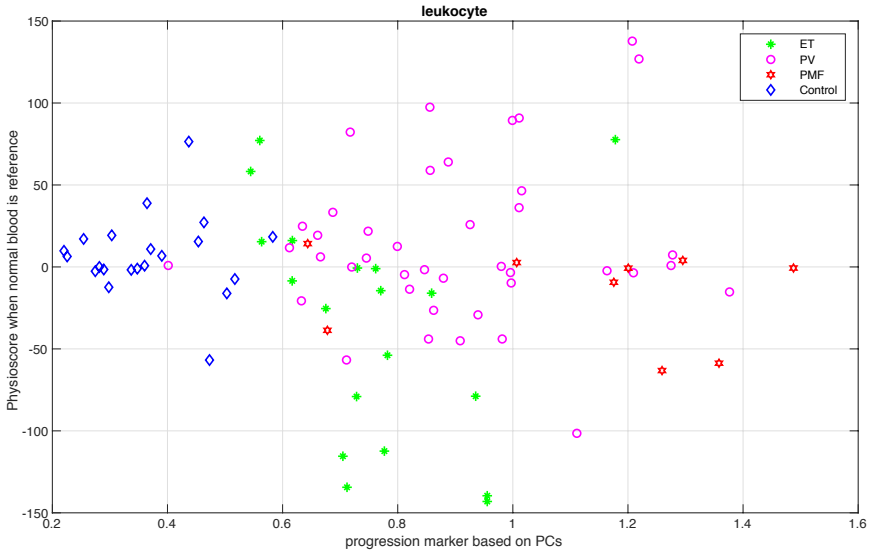
Sfig11: patients' PhysioScores on Monocyte physio axis in a coordinate system with either blood as the reference (upper figure) or HSC as the reference (lower figure). The monotonic decrease of ET samples is observable here only in the upper panel.

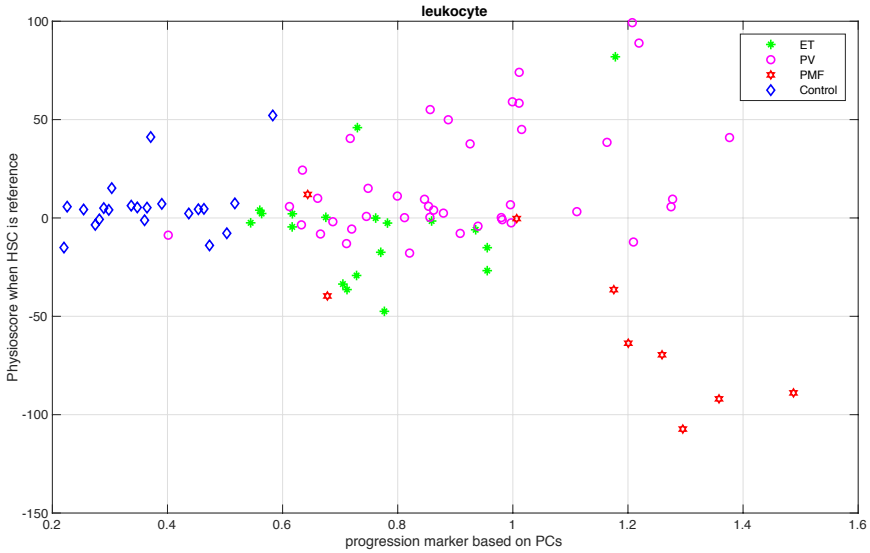


Sfig12: patients' PhysioScores on Lymphocytes physio axis in a coordinate system with either blood as the reference (upper figure) or HSC as the reference (lower figure). The monotonic decrease of ET samples is observable.

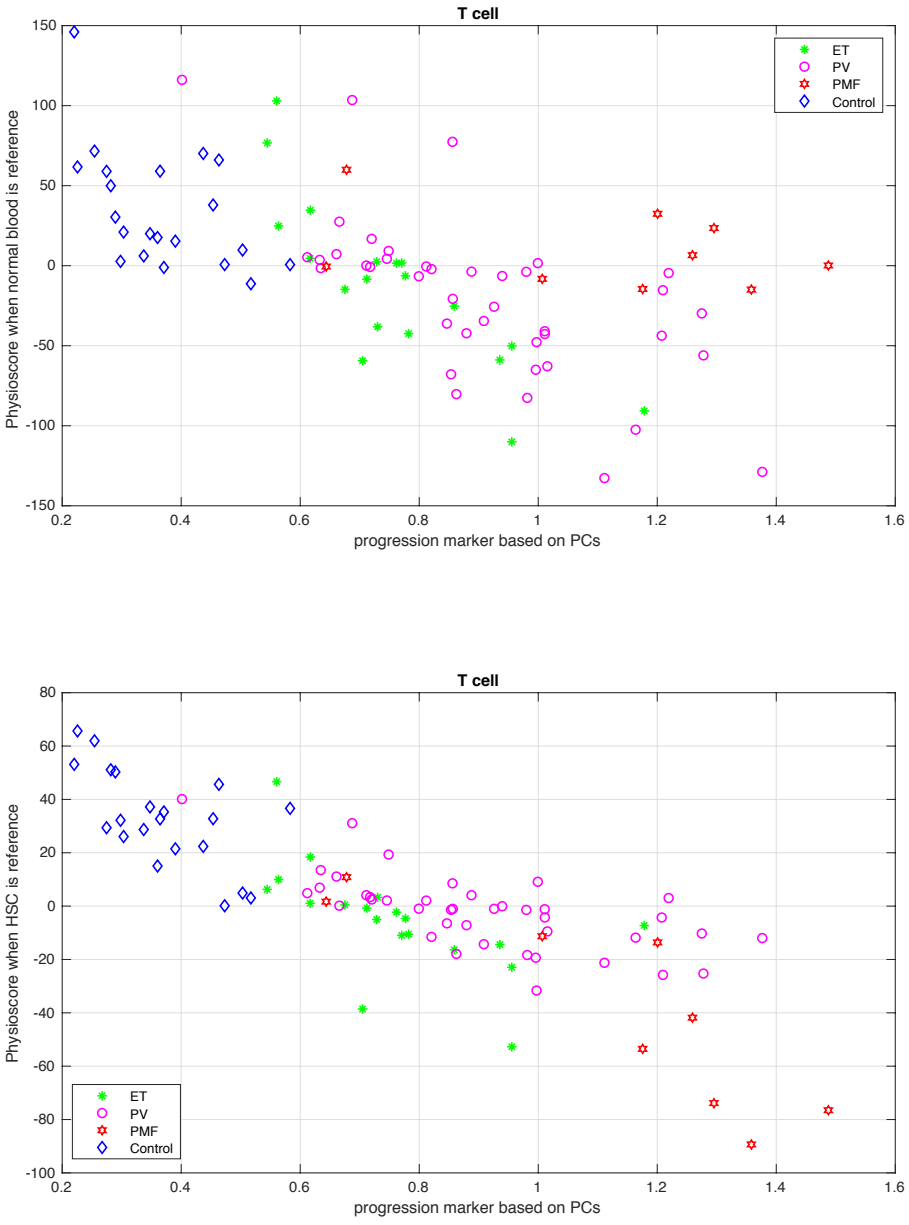


Sfig13: patients' PhysioScores on Erythrocyte physio axis in a coordinate system with either blood as the reference (upper figure) or HSC as the reference (lower figure). The monotonic decrease of ET samples is observable.

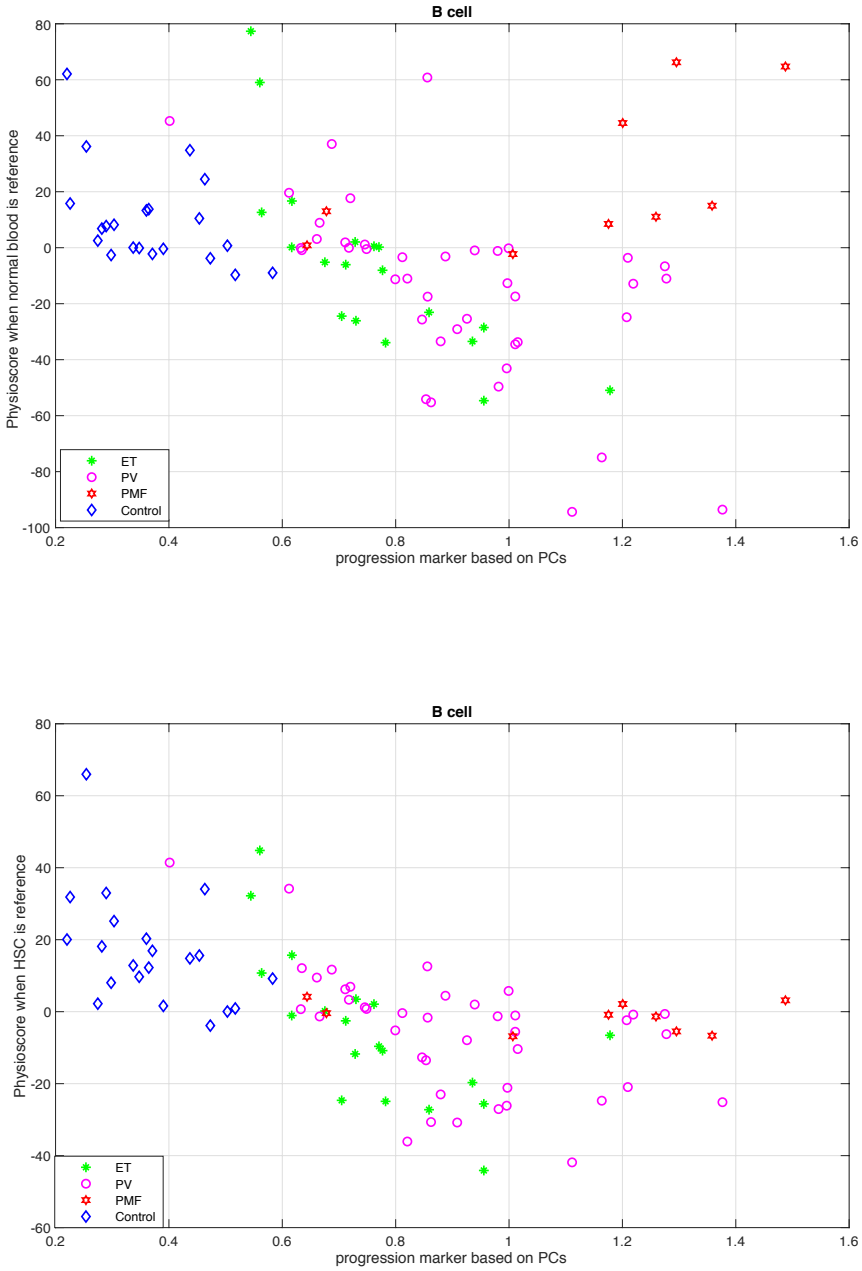




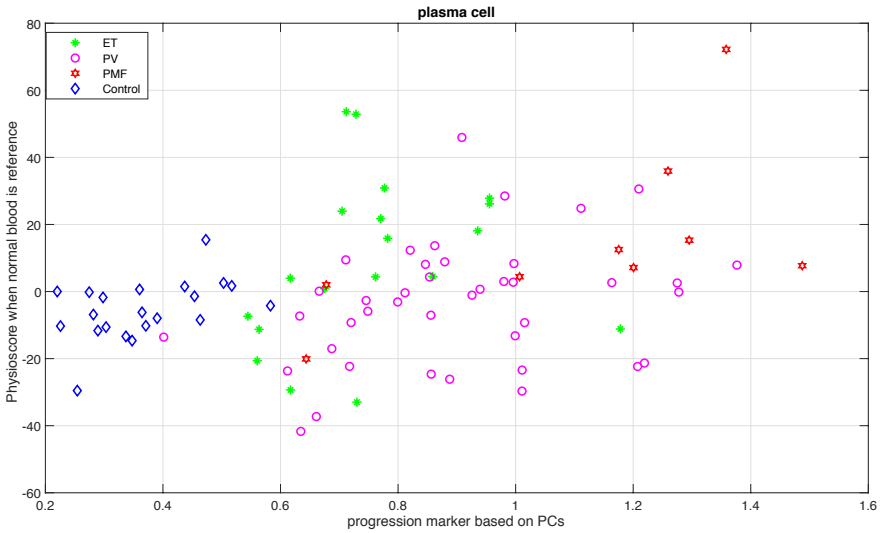
Sfig14: patients' PhysioScores on Leukocyte physio axis in a coordinate system with either blood as the reference (upper figure) or HSC as the reference (lower figure). The monotonic decrease of ET samples is observable only on the upper panel.

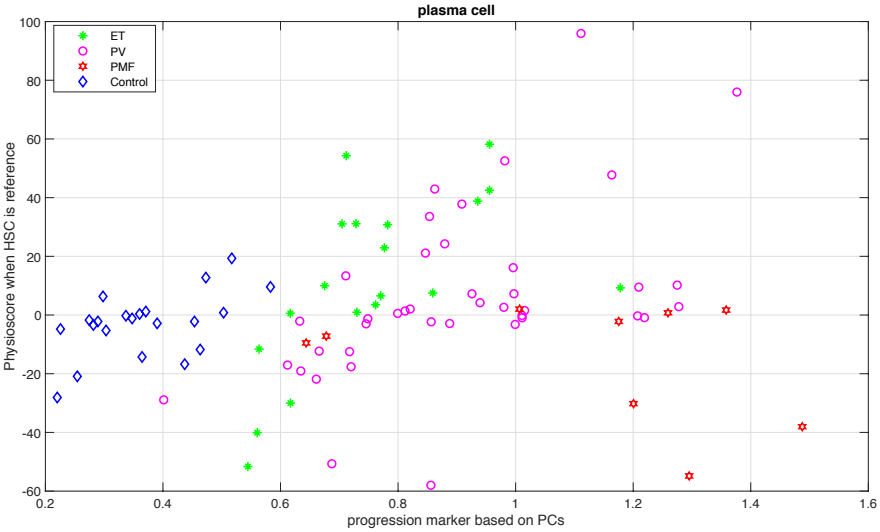


Sfig15: patients' PhysioScores on Tcell physio axis in a coordinate system with either blood as the reference (upper figure) or HSC as the reference (lower figure). The monotonic decrease of ET samples is observable. Also, interestingly Tcell physio scores continuously decrease along the PC-based biomarker starting from control to PMF, when HSC is accounted as the reference of the system.

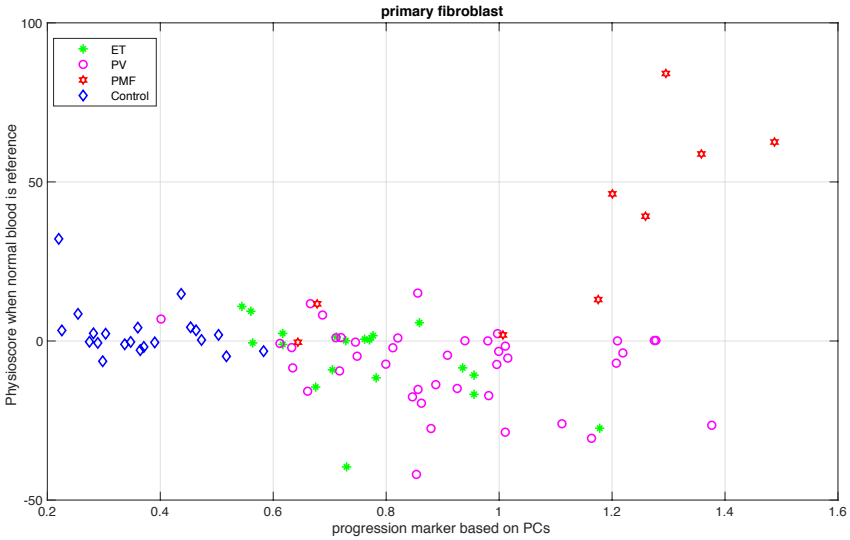


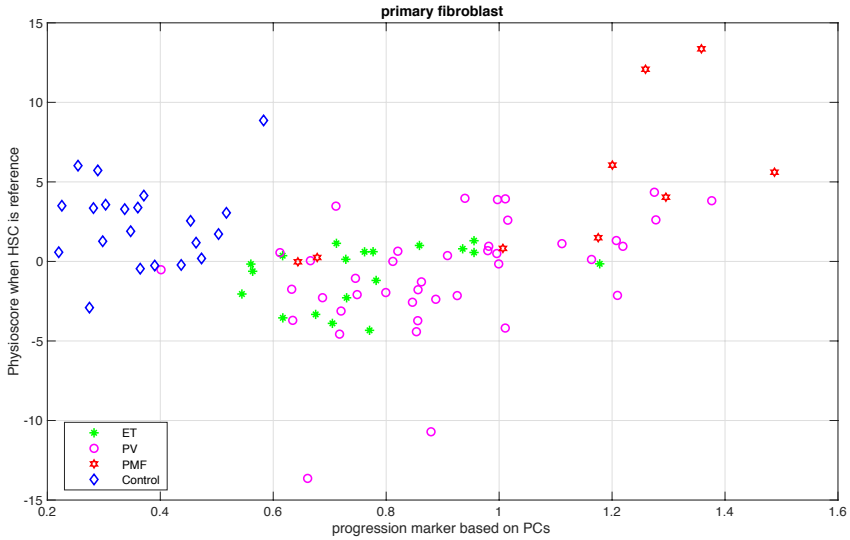
Sfig16: patients' PhysioScores B cell physio axis in a coordinate system with either blood as the reference (upper figure) or HSC as the reference (lower figure). The monotonic decrease of ET samples is observable.





Sfig17: patients' PhysioScores Plasma cell physio axis in a coordinate system with either blood as the reference (upper figure) or HSC as the reference (lower figure). The monotonic increase of ET samples is observable.





Sfig18: patients' PhysioScores primary fibroblast physio axis in a coordinate system with either blood as the reference (upper figure) or HSC as the reference (lower figure). The high positive scores in PMFs support the presence of fibrotic tissue in PMFs.

REFERENCE

1. Fröhlich, H. *et al.* From hype to reality: data science enabling personalized medicine. *BMC Med* **16**, 150 (2018).

2. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).

3. Haeno, H., Levine, R. L., Gilliland, D. G. & Michor, F. A progenitor cell origin of myeloid malignancies. *Proceedings of the National Academy of Sciences* **106**, 16616–16621 (2009).

4. Bernaschi, M., Melchionna, S. & Succi, S. Mesoscopic simulations at the physics-chemistry-biology interface. *Rev. Mod. Phys.* **91**, 025004 (2019).

5. Trefois, C., Antony, P. M. A., Goncalves, J., Skupin, A. & Balling, R. Critical transitions in chronic disease: transferring concepts from ecology to systems medicine. *Curr. Opin. Biotechnol.* **34**, 48–55 (2015).

6. Sorger, P. Quantitative and Systems Pharmacology in the Post-genomic Era: New Approaches to Discovering Drugs and Understanding Therapeutic. 48 (2011).

7. Nowell, P. C. The minute chromosome (Phl) in chronic granulocytic leukemia. *Blut* **8**, 65–66 (1962).

8. A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia identified by Quinacrine Fluorescence and Giemsa Staining | Nature. Available at: https://www.nature.com/articles/243290a0. (Accessed: 5th March 2019)

9. Deininger, M. W. N., Goldman, J. M. & Melo, J. V. The molecular biology of chronic myeloid leukemia. **96**, 15 (2000).

10. Perrotti, D., Jamieson, C., Goldman, J. & Skorski, T. Chronic myeloid leukemia: mechanisms of blastic transformation. *J Clin Invest* **120**, 2254–2264 (2010).

11. Apperley, J. F. Chronic myeloid leukaemia. *The Lancet* **385**, 1447–1459 (2015).

12. SH, S. *et al. WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues*.

13. Jabbour, E. & Kantarjian, H. Chronic myeloid leukemia: 2012 update on diagnosis, monitoring, and management. *Am. J. Hematol.* **87**, 1037–1045 (2012).

14. Granatowicz, A. *et al.* An Overview and Update of Chronic Myeloid Leukemia for Primary Care Physicians. *Korean J Fam Med* **36**, 197–202 (2015).

15. Baccarani, M. et al. European LeukemiaNet recommendations for the management of chronic myeloid leukemia: 2013. Blood 122, 872–884 (2013)

16. Sharma, P. Bone Marrow Histology in CML: A Continuing Relevance. *Journal of Bone Marrow Research* **01**, (2013).

17. Sokal, J. E. *et al.* Prognostic discrimination in 'good-risk' chronic granulocytic leukemia. *Blood* **63**, 789–799 (1984).

18. Hasford, J. *et al.* A New Prognostic Score for Survival of Patients With Chronic Myeloid Leukemia Treated With Interferon AlfaWriting Committee for the Collaborative CML Prognostic Factors Project Group. *J Natl Cancer Inst* **90**, 850–859 (1998).

19. Hasford, J. *et al.* Predicting complete cytogenetic response and subsequent progression-free survival in 2060 patients with CML on imatinib treatment: the EUTOS score. *Blood* **118**, 686–692 (2011).

20. Hochhaus, A. *et al.* Six-year follow-up of patients receiving imatinib for the first-line treatment of chronic myeloid leukemia. *Leukemia* **23**, 1054–1061 (2009).

21. ICLUSIG (ponatinib) tablets. 17

22. Forkner, C. E. & Scott, T. F. M. ARSENIC AS A THERAPEUTIC AGENT IN CHRONIC MYELOGENOUS LEUKEMIA: PRELIMINARY REPORT. *JAMA* **97**, 3–5 (1931).

23. Chronic granulocytic leukaemia: comparison of radiotherapy and busulphan therapy. Report of the Medical Research Council's working party for therapeutic trials in leukaemia. *Br Med J* **1**, 201–208 (1968).

24. Bonifazi, F. *et al.* Chronic myeloid leukemia and interferon-α: a study of complete cytogenetic responders. *Blood* **98**, 3074–3081 (2001).

25. Talpaz, M. *et al.* Hematologic Remission and Cytogenetic Improvement Induced by Recombinant Human Interferon AlphaA in Chronic Myelogenous Leukemia. *New England Journal of Medicine* **314**, 1065–1069 (1986).

26. Hehlmann, R. *et al.* Randomized comparison of interferon-alpha with busulfan and hydroxyurea in chronic myelogenous leukemia. The German CML Study Group. *Blood* **84**, 4064–4077 (1994).

27. Interferon Alfa-2a as Compared with Conventional Chemotherapy for the Treatment of Chronic Myeloid Leukemia. *New England Journal of Medicine* **330**, 820–825 (1994).

28. Allan, N. C., Shepherd, P. C. A. & Richards, S. M. UK Medical Research Council randomised, multicentre trial of interferon-αn1 for chronic myeloid leukaemia: improved survival irrespective of cytogenetic response. *The Lancet* **345**, 1392–1397 (1995).

29. Fefer, A. *et al.* Disappearance of pH1-Positive Cells in Four Patients with Chronic Granulocytic Leukemia after Chemotherapy, Irradiation and Marrow Transplantation from an Identical Twin. *New England Journal of Medicine* **300**, 333–337 (1979).

30. Druker, B. J. *et al.* Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr–Abl positive cells. *Nature Medicine* **2**, 561 (1996).

31. O'Brien, S. G. *et al.* Imatinib Compared with Interferon and Low-Dose Cytarabine for Newly Diagnosed Chronic-Phase Chronic Myeloid Leukemia. *http://dx.doi.org/10.1056/NEJMoa022457* (2009). doi:10.1056/NEJMoa022457

32. Vainchenker, W. & Kralovics, R. Genetic basis and molecular pathophysiology of classical myeloproliferative neoplasms. *Blood* **129**, 667–679 (2017).

33. Gertz, M. A. Waldenström Macroglobulinemia: 2012 update on diagnosis, risk stratification, and management. *American Journal of Hematology* **87**, 503–510 (2012).

34. Moulard, O. *et al.* Epidemiology of myelofibrosis, essential thrombocythemia, and polycythemia vera in the European Union. *European Journal of Haematology* **92**, 289–297 (2014).

35. Srour, S. A. *et al.* Incidence and patient survival of myeloproliferative neoplasms and myelodysplastic/myeloproliferative neoplasms in the United States, 2001-2012. *Br J Haematol* **174**, 382–396 (2016).

36. Hultcrantz, M. *et al.* Patterns of Survival Among Patients With Myeloproliferative Neoplasms Diagnosed in Sweden From 1973 to 2008: A Population-Based Study. *J Clin Oncol* **30**, 2995–3001 (2012).

37. Tefferi, A. & Vainchenker, W. Myeloproliferative Neoplasms: Molecular Pathophysiology, Essential Clinical Understanding, and Treatment Strategies. *Journal of Clinical Oncology* **29**, 573–582 (2011).

38. James, C. *et al.* A unique clonal *JAK2* mutation leading to constitutive signalling causes polycythaemia vera. *Nature* **434**, 1144–1148 (2005).

39. Kralovics, R. *et al.* A Gain-of-Function Mutation of JAK2 in Myeloproliferative Disorders. *http://dx.doi.org/10.1056/NEJMoa051113* (2009). doi:10.1056/NEJMoa051113

40. Levine, R. L. *et al.* Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. *Cancer Cell* **7**, 387–397 (2005).

41. Baxter, E. J. *et al.* Acquired mutation of the tyrosine kinase JAK2 in human myeloproliferative disorders. *The Lancet* **365**, 1054–1061 (2005).

42. Scott, L. M. *et al.* JAK2 Exon 12 Mutations in Polycythemia Vera and Idiopathic Erythrocytosis. *The New England Journal of Medicine* 10 (2007).

43. Pikman, Y. *et al.* MPLW515L Is a Novel Somatic Activating Mutation in Myelofibrosis with Myeloid Metaplasia. *PLoS Med* **3**, (2006).

44. Pardanani, A. D. *et al.* MPL515 mutations in myeloproliferative and other myeloid disorders: a study of 1182 patients. *Blood* **108**, 3472–3476 (2006).

45. Bilbao-Sieyro, C., Florido, Y. & Gómez-Casares, M. T. CALR mutation characterization in myeloproliferative neoplasms. *Oncotarget* **7**, 52614–52617 (2016).

46. Hookham, M. B. *et al.* The myeloproliferative disorder–associated JAK2 V617F mutant escapes negative regulation by suppressor of cytokine signaling 3. *Blood* **109**, 4924–4929 (2007).

47. Spivak, J. L. The chronic myeloproliferative disorders: Clonality and clinical heterogeneity. *Seminars in Hematology* **41**, 1–5 (2004).

48. Arber, D. A. *et al.* The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391–2405 (2016).

49. Barbui, T. *et al.* Survival and disease progression in essential thrombocythemia are significantly influenced by accurate morphologic diagnosis: an international study. *J. Clin. Oncol.* **29**, 3179–3184 (2011).

50. Kralovics, R., Guan, Y. & Prchal, J. T. Acquired uniparental disomy of chromosome 9p is a frequent stem cell defect in polycythemia vera. *Experimental Hematology* **30**, 229–236 (2002).

51. Rumi, E. *et al.* JAK2 or CALR mutation status defines subtypes of essential thrombocythemia with substantially different clinical course and outcomes. *Blood* **123**, 1544–1551 (2014).

52. Vannucchi, A. M., Guglielmelli, P. & Tefferi, A. Advances in Understanding and Management of Myeloproliferative Neoplasms. *CA: A Cancer Journal for Clinicians* **59**, 171–191 (2009).

53. Passamonti, F. *et al.* Relation between JAK2 (V617F) mutation status, granulocyte activation, and constitutive mobilization of CD34+ cells into peripheral blood in myeloproliferative disorders. *Blood* **107**, 3676–3682 (2006).

54. Vannucchi, A. M., Antonioli, E., Guglielmelli, P., Pardanani, A. & Tefferi, A. Clinical correlates of JAK2V617F presence or allele burden in myeloproliferative neoplasms: a critical reappraisal. *Leukemia* **22**, 1299–1307 (2008).

55. Shammo, J. M. & Stein, B. L. Mutations in MPNs: prognostic implications, window to biology, and impact on treatment decisions. *Hematology* **2016**, 552–560 (2016).

56. Akpınar, T. S., Hançer, V. S., Nalçacı, M. & Diz-Küçükkaya, R. MPL W515L/K Mutations in Chronic Myeloproliferative Neoplasms. *Turk J Haematol* **30**, 8–12 (2013).

57. Klampfl, T. *et al.* Somatic mutations of calreticulin in myeloproliferative neoplasms. *N. Engl. J. Med.* **369**, 2379–2390 (2013).

58. Nangalia, J. *et al.* Somatic CALR mutations in myeloproliferative neoplasms with nonmutated JAK2. *N. Engl. J. Med.* **369**, 2391–2405 (2013).

59. Cavalloni, C. *et al.* Sequential evaluation of CALR mutant burden in patients with myeloproliferative neoplasms. *Oncotarget* **8**, 33416–33421 (2017).

60. Oh, S. T. *et al.* Novel mutations in the inhibitory adaptor protein LNK drive JAK-STAT signaling in patients with myeloproliferative neoplasms. *Blood* **116**, 988–992 (2010).

61. McMullin, M. F. & Cario, H. LNK mutations and myeloproliferative disorders. *American Journal of Hematology* **91**, 248–251 (2016).

62. Pardanani, A. *et al.* LNK mutation studies in blast-phase myeloproliferative neoplasms, and in chronic-phase disease with *TET2*, *IDH*, *JAK2* or *MPL* mutations. *Leukemia* **24**, 1713–1718 (2010).

63. Marty, C. *et al.* Calreticulin mutants in mice induce an MPL-dependent thrombocytosis with frequent progression to myelofibrosis. *Blood* **127**, 1317–1324 (2016).

64. Balligand, T. *et al.* Crispr/Cas9 Engineered 61bp Deletion in the Calr Gene of Mice Leads to Development of Thrombocytosis. 3

65. Pasquier, F., Cabagnols, X., Secardin, L., Plo, I. & Vainchenker, W. Myeloproliferative Neoplasms: JAK2 Signaling Pathway as a Central Target for Therapy. *Clinical Lymphoma Myeloma and Leukemia* **14**, S23–S35 (2014).

66. Stegelmann, F. *et al.* High-resolution single-nucleotide polymorphism array-profiling in myeloproliferative neoplasms identifies novel genomic aberrations. *Haematologica* **95**, 666–669 (2010).

67. Rampal, R. *et al.* Genomic and functional analysis of leukemic transformation of myeloproliferative neoplasms. *PNAS* **111**, E5401–E5410 (2014).

68. Cazzola, M. Introduction to a review series: the 2016 revision of the WHO classification of tumors of hematopoietic and lymphoid tissues. *Blood* **127**, 2361–2364 (2016).

69. Schafer, A. I. Thrombocytosis. *New England Journal of Medicine* **350**, 1211–1219 (2004).

70. Rumi, E. & Cazzola, M. How I treat essential thrombocythemia. **128**, 13 (2016).

71. Pearson, T. C. *et al.* Interpretation of measured red cell mass and plasma volume in adults: Expert Panel on Radionuclides of the International Council for Standardization in Haematology. *Br. J. Haematol.* **89**, 748–756 (1995).

72. Rumi, E. & Cazzola, M. Diagnosis, risk stratification, and response evaluation in classical myeloproliferative neoplasms. *Blood* **129**, 680–692 (2017).

73. Cervantes, F. How I treat myelofibrosis. *Blood* **124**, 2635–2642 (2014).

74. Essential thrombocythemia: A review of diagnostic and pathologic features | Request PDF. *ResearchGate* Available at: https://www.researchgate.net/publication/6907663_Essential_thrombocythemia_A_review_of_diagnostic_and_pathologic_features. (Accessed: 6th March 2019)

75. Lakey, M. A. *et al.* Bone Marrow Morphologic Features in Polycythemia Vera With JAK2 Exon 12 Mutations. *Am J Clin Pathol* **133**, 942–948 (2010).

76. Pozdnyakova, O., Hasserjian, R. P., Verstovsek, S. & Orazi, A. Impact of Bone Marrow Pathology on the Clinical Management of Philadelphia Chromosome–Negative Myeloproliferative Neoplasms. *Clinical Lymphoma Myeloma and Leukemia* **15**, 253–261 (2015).

77. Vannucchi, A. M. & Harrison, C. N. Emerging treatments for classical myeloproliferative neoplasms. *Blood* **129**, 693–703 (2017).

78. Barbui, T. *et al.* Philadelphia-negative classical myeloproliferative neoplasms: critical concepts and management recommendations from European LeukemiaNet. *J. Clin. Oncol.* **29**, 761–770 (2011).

79. Barbui, T. *et al.* Development and validation of an International Prognostic Score of thrombosis in World Health Organization-essential thrombocythemia (IPSET-thrombosis). *Blood* **120**, 5128–5133; quiz 5252 (2012).

80. Passamonti, F. *et al.* A prognostic model to predict survival in 867 World Health Organization-defined essential thrombocythemia at diagnosis: a study by the International Working Group on Myelofibrosis Research and Treatment. *Blood* **120**, 1197–1201 (2012).

81. Rumi, E. & Cazzola, M. Diagnosis, risk stratification, and response evaluation in classical myeloproliferative neoplasms. *Blood* **129**, 680–692 (2017).

82. Tefferi, A. *et al.* Survival and prognosis among 1545 patients with contemporary polycythemia vera: an international study. *Leukemia* **27**, 1874–1881 (2013).

83. Cervantes, F. *et al.* New prognostic scoring system for primary myelofibrosis based on a study of the International Working Group for Myelofibrosis Research and Treatment. *Blood* **113**, 2895–2901 (2009).

84. Passamonti, F. *et al.* A dynamic prognostic model to predict survival in primary myelofibrosis: a study by the IWG-MRT (International Working Group for Myeloproliferative Neoplasms Research and Treatment). *Blood* **115**, 1703–1708 (2010).

85. Gangat, N. *et al.* DIPSS plus: a refined Dynamic International Prognostic Scoring System for primary myelofibrosis that incorporates prognostic information from karyotype, platelet count, and transfusion status. *J. Clin. Oncol.* **29**, 392–397 (2011).

86. Chan, D. & Koren-Michowitz, M. Update on JAK2 Inhibitors in Myeloproliferative Neoplasm. *Ther Adv Hematol* **2**, 61–71 (2011).

87. Martínez-Trillos, A. *et al.* Efficacy and tolerability of hydroxyurea in the treatment of the hyperproliferative manifestations of myelofibrosis: results in 40 patients. *Ann. Hematol.* **89**, 1233–1237 (2010).

88. de Melo Campos, P. Primary myelofibrosis: current therapeutic options. *Rev Bras Hematol Hemoter* **38**, 257–263 (2016).

89. Bose, P. & Verstovsek, S. JAK2 inhibitors for myeloproliferative neoplasms: what is next? *Blood* **130**, 115–125 (2017).

90. Verstovsek, S. *et al.* A Double-Blind, Placebo-Controlled Trial of Ruxolitinib for Myelofibrosis. *New England Journal of Medicine* **366**, 799–807 (2012).

91. Harrison, C. *et al.* JAK inhibition with ruxolitinib versus best available therapy for myelofibrosis. *N. Engl. J. Med.* **366**, 787–798 (2012).

92. Verstovsek, S. *et al.* Ruxolitinib versus best available therapy in patients with polycythemia vera: 80-week follow-up from the RESPONSE trial. *Haematologica* **101**, 821–829 (2016).

93. Mesa, R. *et al.* Changes in quality of life and disease-related symptoms in patients with polycythemia vera receiving ruxolitinib or standard therapy. *Eur. J. Haematol.* **97**, 192–200 (2016).

94. Vannucchi, A. M. *et al.* Ruxolitinib versus standard therapy for the treatment of polycythemia vera. *N. Engl. J. Med.* **372**, 426–435 (2015).

95. Pieri, L. *et al.* JAK2V617F complete molecular remission in polycythemia vera/essential thrombocythemia patients treated with ruxolitinib. *Blood* **125**, 3352–3353 (2015).

96. Verstovsek, S. *et al.* Long-Term Results from a Phase II Open-Label Study of Ruxolitinib in Patients with Essential Thrombocythemia Refractory to or Intolerant of Hydroxyurea. *Blood* **124**, 1847–1847 (2014).

97. Pardanani, A. *et al.* Safety and Efficacy of Fedratinib in Patients With Primary or Secondary Myelofibrosis: A Randomized Clinical Trial. *JAMA Oncol* **1**, 643–651 (2015).

98. Verstovsek, S. *et al.* A phase I, open-label, dose-escalation, multicenter study of the JAK2 inhibitor NS-018 in patients with myelofibrosis. *Leukemia* **31**, 393–402 (2017).

99. Mascarenhas, J. O. *et al.* Primary analysis of a phase II open-label trial of INCB039110, a selective JAK1 inhibitor, in patients with myelofibrosis. *Haematologica* **102**, 327–335 (2017).

100. Mesa, R. A. *et al.* Pacritinib versus best available therapy for the treatment of myelofibrosis irrespective of baseline cytopenias (PERSIST-1): an international, randomised, phase 3 trial. *Lancet Haematol* **4**, e225–e236 (2017).

101. Gupta, V. *et al.* A phase 1/2, open-label study evaluating twice-daily administration of momelotinib in myelofibrosis. *Haematologica* **102**, 94–102 (2017).

102. Pardanani, A. *et al.* Phase I/II Study of CYT387, a JAK1/JAK2 Inhibitor for the Treatment of Myelofibrosis. *Blood* **120**, 178–178 (2012).

103. Klampfl, T. *et al.* Genome integrity of myeloproliferative neoplasms in chronic phase and during disease progression. *Blood* **118**, 167–176 (2011).

104. Campbell, P. J. & Green, A. R. The myeloproliferative disorders. *N. Engl. J. Med.* **355**, 2452–2466 (2006).

105. Tam, C. S. *et al.* Dynamic Model for Predicting Death Within 12 Months in Patients With Primary or Post–Polycythemia Vera/Essential Thrombocythemia Myelofibrosis. *J Clin Oncol* **27**, 5587–5593 (2009).

106. Abdulkarim, K. *et al.* AML transformation in 56 patients with Ph− MPD in two well defined populations. *European Journal of Haematology* **82**, 106–111 (2009).

107. Yogarajah, M. & Tefferi, A. Leukemic Transformation in Myeloproliferative Neoplasms. *Mayo Clinic Proceedings* **92**, 1118–1128 (2017).

108. Lundberg, P. *et al.* Clonal evolution and clinical correlates of somatic mutations in myeloproliferative neoplasms. *Blood* **123**, 2220–2228 (2014).

109. Larsen, T. S., Pallisgaard, N., Møller, M. B. & Hasselbalch, H. C. The JAK2 V617F allele burden in essential thrombocythemia, polycythemia vera and primary myelofibrosis – impact on disease phenotype. *European Journal of Haematology* **79**, 508–515 (2007).

110. Hasselbalch, H. C. Chronic inflammation as a promotor of mutagenesis in essential thrombocythemia, polycythemia vera and myelofibrosis. A human inflammation model for cancer development? *Leukemia Research* **37**, 214–220 (2013).

111. Hasselbalch, H. C. & Bjørn, M. E. MPNs as Inflammatory Diseases: The Evidence, Consequences, and Perspectives. *Mediators of Inflammation* (2015). doi:10.1155/2015/102476

112. Hemminki, A. & Hemminki, K. The Genetic Basis of Cancer. in *Cancer Gene Therapy* (eds. Curiel, D. T. & Douglas, J. T.) 9–18 (Humana Press, 2005). doi:10.1007/978-1-59259-785-7_2

113. Uzman, A. The biology of cancer by R. A. Weinberg. *Biochemistry and Molecular Biology Education* **36**, 320–321 (2008).

114. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100**, 57–70 (2000).

115. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).

116. Nunney, L. Lineage selection and the evolution of multistage carcinogenesis. *Proc Biol Sci* **266**, 493–498 (1999).

117. Davies, P. C., Demetrius, L. & Tuszynski, J. A. Cancer as a dynamical phase transition. *Theor Biol Med Model* **8**, 30 (2011).

118. Altrock, P. M., Liu, L. L. & Michor, F. The mathematics of cancer: integrating quantitative models. *Nature Reviews Cancer* **15**, 730–745 (2015).

119. Michor, F. *et al.* Dynamics of chronic myeloid leukaemia. *Nature* **435**, 1267–1270 (2005).

120. Tang, M. *et al.* Dynamics of chronic myeloid leukemia response to long-term targeted therapy reveal treatment effects on leukemic stem cells. *Blood* **118**, 1622–1631 (2011).

121. Dingli, D., Traulsen, A., Lenaerts, T. & Pacheco, J. M. Evolutionary dynamics of chronic myeloid leukemia. *Genes Cancer* **1**, 309–315 (2010).

122. Mogk, G., Mrziglod, Th. & Schuppert, A. Application of Hybrid Models in Chemical Industry. in *Computer Aided Chemical Engineering* (eds. Grievink, J. & van Schijndel, J.) **10**, 931–936 (Elsevier, 2002).

123. Fokas, A. S., Keller, J. B. & Clarkson, B. D. Mathematical model of granulocytopoiesis and chronic myelogenous leukemia. *Cancer Res.* **51**, 2084–2091 (1991).

124. Moore, H. & Li, N. K. A mathematical model for chronic myelogenous leukemia (CML) and T cell interaction. *Journal of Theoretical Biology* **227**, 513–523 (2004).

125. Dingli, D., Traulsen, A. & Pacheco, J. M. Chronic Myeloid Leukemia: Origin, Development, Response to Therapy, and Relapse. *Clinical Leukemia* **2**, 133–139 (2008).

126. Dingli, D., Traulsen, A. & Pacheco, J. M. Compartmental Architecture and Dynamics of Hematopoiesis. *PLOS ONE* **2**, e345 (2007).

127. Adimy, M., Crauste, F. & Ruan, S. A Mathematical Study of the Hematopoiesis Process with Applications to Chronic Myelogenous Leukemia. *SIAM Journal on Applied Mathematics* **65**, 1328–1352 (2005).

128. Wodarz, D. Stem cell regulation and the development of blast crisis in chronic myeloid leukemia: Implications for the outcome of Imatinib treatment and discontinuation. *Medical Hypotheses* **70**, 128–136 (2008).

129. Zhang, B., Tanaka, G., Chen, L. & Aihara, K. Dynamical modeling of chronic myeloid leukemia progression and the development of mutations. in *2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops* 130–135 (2012). doi:10.1109/BIBMW.2012.6470294

130. Flå, T., Rupp, F. & Woywod, C. Bifurcation patterns in generalized models for the dynamics of normal and leukemic stem cells with signaling. *Mathematical Methods in the Applied Sciences* **38**, 3392–3407 (2015).

131. Ledzewicz, U. & Moore, H. Dynamical Systems Properties of a Mathematical Model for the Treatment of CML. *Applied Sciences* **6**, 291 (2016).

132. Woywod, C., Gruber, F. X., Engh, R. A. & Flå, T. Dynamical models of mutated chronic myelogenous leukemia cells for a post-imatinib treatment scenario: Response to dasatinib or nilotinib therapy. *PLOS ONE* **12**, e0179700 (2017).

133. Sasaki, K. *et al.* Clinical Application of Artificial Intelligence in Patients with Chronic Myeloid Leukemia in Chronic Phase. *Blood* **128**, 940–940 (2016).

134. Khosla, E. & Ramesh, D. Phase classification of chronic myeloid leukemia using convolution neural networks. in *2018 4th International Conference on Recent Advances in Information Technology (RAIT)* 1–6 (2018). doi:10.1109/RAIT.2018.8389068

135. Dincer, A. B., Celik, S., Hiranuma, N. & Lee, S.-I. DeepProfile: Deep learning of cancer molecular profiles for precision medicine. *bioRxiv* (2018). doi:10.1101/278739

136. Wang, X. *et al.* Automated identification of abnormal metaphase chromosome cells for the detection of chronic myeloid leukemia using microscopic images. *JBO* **15**, 046026 (2010).

137. Yeung, K. Y. *et al.* Predicting relapse prior to transplantation in chronic myeloid leukemia by integrating expert knowledge and expression data. *Bioinformatics* **28**, 823–830 (2012).

138. Oehler, V. G. *et al.* The derivation of diagnostic markers of chronic myeloid leukemia progression from microarray data. *Blood* **114**, 3292–3298 (2009).

139. Attolini, C. S.-O. *et al.* A mathematical framework to determine the temporal sequence of somatic genetic events in cancer. *Proceedings of the National Academy of Sciences* **107**, 17604–17609 (2010).

140. Andersen, M. *et al.* Mathematical modelling as a proof of concept for MPNs as a human inflammation model for cancer development. *PLOS ONE* **12**, e0183620 (2017).

141. Dingli, D. & Michor, F. Successful Therapy Must Eradicate Cancer Stem Cells. *STEM CELLS* **24**, 2603–2610 (2006).

142. Fiedler, B. & Schuppert, A. Local identification of hybrid models with tree structure. *IMA Journal of Applied Mathematics* **73**, (2008).

143. Dakos, V. *et al.* Methods for Detecting Early Warnings of Critical Transitions in Time Series Illustrated Using Simulated Ecological Data. *PLOS ONE* **7**, e41010 (2012).

144. Entropy as a method to investigate complex biological systems. An alternative view on the biological transition from healthy aging to frailty | Geriatric Care. Available at: https://www.pagepressjournals.org/index.php/gc/article/view/6755/6776#content/citation_reference_10. (Accessed: 31st May 2019)

145. Brandani, G. B. *et al.* Quantifying Disorder through Conditional Entropy: An Application to Fluid Mixing. *PLoS ONE* **8**, e65617 (2013).

146. Neural Networks M. Hajek 2005Neural Networks. Available at: http://ai4trade.com/GeneticAlgorithmsInForex/neural-networks-m-hajek-2005. (Accessed: 29th May 2019)

147. Pagariya, R. & Bartere, M. Review Paper on Artificial Neural Networks. *International Journal of Advanced Research in Computer Science* **4**, (2013).

148. Gerven, M. van & Bohte, S. *Artificial Neural Networks as Models of Neural Information Processing*. (Frontiers Media SA, 2018).

149. Neural Networks. Available at: https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html. (Accessed: 3rd June 2019)

150. Tu, J. V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology* **49**, 1225–1231 (1996).

151. Zayegh, A. & Bassam, N. A. Neural Network Principles and Applications. *Digital Systems* (2018). doi:10.5772/intechopen.80416

152. Papik, K. *et al.* Application of neural networks in medicine - A review. in (1998).

153. Anrew Ng. *Machine learning course on course area*.

154. Menard, S. W. *Applied logistic regression analysis*. (Sage Publications, 1995).

155. Boser, B. E., Guyon, I. M. & Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* 144–152 (ACM, 1992). doi:10.1145/130385.130401

156. Noble, W. S. What is a support vector machine? *Nature Biotechnology* **24**, 1565 (2006).

157. Goel, E. & Abhilasha, E. Random Forest: A Review. in (2017). doi:10.23956/ijarcsse/V7I1/01113

158. Anuradha & Gupta, G. A self explanatory review of decision tree classifiers. in *International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014)* 1–7 (2014). doi:10.1109/ICRAIE.2014.6909245

159. SONG, Y. & LU, Y. Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry* **27**, 130–135 (2015).

160. Biau, G. & Scornet, E. A random forest guided tour. *TEST* **25**, 197–227 (2016).

161. Aslam, J. A., Popa, R. A. & Rivest, R. L. On Estimating the Size and Confidence of a Statistical Audit. 12

162. Cutler, F. original by L. B. and A. & Wiener, R. port by A. L. and M. *randomForest: Breiman and Cutler's Random Forests for Classification and Regression*. (2018).

163. Vanwinckelen, G. & Blockeel, H. On Estimating Model Accuracy with Repeated Cross-Validation. 6

164. Arlot, S. & Celisse, A. A survey of cross-validation procedures for model selection. *Statist. Surv.* **4**, 40–79 (2010).

165. F.R.S, K. P. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**, 559–572 (1901).

166. Einasto, M. *et al.* SDSS DR7 superclusters. Principal component analysis. *A&A* **535**, A36 (2011).

167. Lenz, M., Schuldt, B. M., Müller, F.-J. & Schuppert, A. PhysioSpace: Relating Gene Expression Experiments from Heterogeneous Sources Using Shared Physiological Processes. *PLoS ONE* **8**, e77627 (2013).

168. Radich, J. P. *et al.* Gene expression changes associated with progression and response in chronic myeloid leukemia. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 2794–2799 (2006).

169. Skov, V. *et al.* Whole-blood transcriptional profiling of interferon-inducible genes identifies highly upregulated IFI27 in primary myelofibrosis. *Eur. J. Haematol.* **87**, 54–60 (2011).

170. GSG-MPN. Available at: https://www.cto-im3.de/gsgmpn/. (Accessed: 5th June 2019)

171. Schneckener, S., Arden, N. S. & Schuppert, A. Quantifying stability in gene list ranking across microarray derived clinical biomarkers. *BMC Med Genomics* **4**, 73 (2011).

172. Haibe-Kains, B. *et al.* Inconsistency in large pharmacogenomic studies. *Nature* **504**, 389–393 (2013).

173. Cramer-Morales, K. *et al.* Personalized synthetic lethality induced by targeting RAD52 in leukemias identified by gene mutation and expression profile. *Blood* **122**, 1293–1304 (2013).

174. Principal component analysis of raw data - MATLAB pca - MathWorks Deutschland. Available at: https://de.mathworks.com/help/stats/pca.html. (Accessed: 7th March 2019)

175. Brehme, M. *et al.* Combined Population Dynamics and Entropy Modelling Supports Patient Stratification in Chronic Myeloid Leukemia. *Scientific Reports* **6**, (2016).

176. Gordon, M. Y. *et al.* Cell biology of CML cells. *Leukemia* **13 Suppl 1**, S65-71 (1999).

177. Shet, A., Jahagirdar, B. N. & Verfaillie, C. M. Chronic myelogenous leukemia: mechanisms underlying disease progression. *Leukemia* **16**, 1402–1411 (2002).

178. Berretta, R. & Moscato, P. Cancer Biomarker Discovery: The Entropic Hallmark. *PLOS ONE* **5**, e12262 (2010).

179. Banerji, C. R. S., Severini, S., Caldas, C. & Teschendorff, A. E. Intra-Tumour Signalling Entropy Determines Clinical Outcome in Breast and Lung Cancer. *PLOS Computational Biology* **11**, e1004115 (2015).

180. Choose an ODE Solver - MATLAB & Simulink - MathWorks Deutschland. Available at: https://de.mathworks.com/help/matlab/math/choose-an-ode-solver.html. (Accessed: 8th March 2019)

181. Create linear regression model - MATLAB fitlm - MathWorks Deutschland. Available at: https://de.mathworks.com/help/stats/fitlm.html. (Accessed: 7th March 2019)

182. PANTHER - Gene List Analysis. Available at: http://www.pantherdb.org/. (Accessed: 8th March 2019)

183. Haken, H. *Synergetics: An Introduction Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry and Biology*. (Springer-Verlag, 1978).

184. Skov, V. *et al.* Whole-blood transcriptional profiling of interferon-inducible genes identifies highly upregulated IFI27 in primary myelofibrosis: Transcriptional profiling of IFN-inducible genes in CMPNs. *European Journal of Haematology* **87**, 54–60 (2011).

185. Barbui, T. *et al.* The 2016 WHO classification and diagnostic criteria for myeloproliferative neoplasms: document summary and in-depth discussion. *Blood Cancer J* **8**, (2018).

186. GEO Accession viewer. Available at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE120362. (Accessed: 12th March 2019)

187. org.Hs.eg.db. *Bioconductor* Available at: http://bioconductor.org/packages/org.Hs.eg.db/. (Accessed: 11th March 2019)

188. Gaujoux, R. & Seoighe, C. CellMix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics* **29**, 2211–2212 (2013).

189. Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data | Genes & Immunity. Available at: https://www.nature.com/articles/6364173. (Accessed: 22nd March 2019)

190. Murtagh, F. & Legendre, P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification* **31**, 274–295 (2014).

191. Hartigan, J. A. & Wong, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics* **28**, 100 (1979).

192. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53–65 (1987).

193. Biobase. *Bioconductor* Available at: http://bioconductor.org/packages/Biobase/. (Accessed: 8th March 2019)

194. GEOquery. *Bioconductor* Available at: http://bioconductor.org/packages/GEOquery/. (Accessed: 8th March 2019)

195. Standardized z-scores - MATLAB zscore - MathWorks Deutschland. Available at: https://de.mathworks.com/help/stats/zscore.html. (Accessed: 8th March 2019)

196. *Training of Neural Networks. Contribute to bips-hb/neuralnet development by creating an account on GitHub*. (BIPS, 2019).

197. Press, T. M. Adaptive Computation and Machine Learning series. *The MIT Press* Available at: https://mitpress.mit.edu/books/series/adaptive-computation-and-machine-learning-series. (Accessed: 11th March 2019)

198. ggplot2 package | R Documentation. Available at: https://www.rdocumentation.org/packages/ggplot2/versions/3.1.0. (Accessed: 11th March 2019)

199. Ripley, B. & Venables, W. *nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models*. (2016).

200. Friedman, J. *et al. glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*. (2018).

201. Steinwart, I. & Thomann, P. *liquidSVM: A Fast and Versatile SVM Package*. (2019).

202. Lukk, M. *et al.* A global map of human gene expression. *Nat Biotechnol* **28**, 322–324 (2010).

203. mi package | R Documentation. Available at: https://www.rdocumentation.org/packages/mi/versions/1.0. (Accessed: 8th April 2019)

204. Lammers, B. *ANN2: Artificial Neural Networks for Anomaly Detection*. (2019).

205. Zhao, Y. & Simon, R. Gene expression deconvolution in clinical samples. *Genome Med* **2**, 93 (2010).

206. Abbas, A. R. *et al.* Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. *Genes and Immunity* **6**, 319–331 (2005).

207. Bioconductor - Home. Available at: https://www.bioconductor.org/. (Accessed: 22nd March 2019)

208. BiocInstaller. *Bioconductor* Available at: http://bioconductor.org/packages/BiocInstaller/. (Accessed: 22nd March 2019)