Models of living tissues, numerical simulations and immunotherapy of cancers

Alexandre Poulain

Sorbonne Université
LJLL

École doctorale **École Doctorale Sciences Mathématiques de Paris Centre**

Unité de recherche **Laboratoire Jacques-Louis Lions**

Thèse présentée par **Poulain Alexandre**

Soutenue le **27 septembre 2021**

En vue de l'obtention du grade de docteur de Sorbonne Université

Discipline **Mathématiques appliquées**

Spécialité **Analyse numérique**

Modèles de tissus vivants, simulations numériques et immunothérapie des cancers

Thèse dirigée par	Benoît Perthame	directeur
	Tommaso Lorenzi	co-directeur

Composition du jury

Rapporteurs	Alain Miranville	Université de Poitiers
	Christian Schmeiser	University of Vienna
Examinateurs	Marie Doumic	INRIA
	Katharina Schratz	Sorbonne Université
	Florence Hubert	Aix-Marseille Université
	Pasquale Ciarletta	Politecnico di Milano
Directeurs de thèse	Benoît Perthame	Sorbonne Université
	Tommaso Lorenzi	Politecnico di Torino

Mots clés : Modèles de tissus vivants, Analyse numérique, Équation de Cahn-Hilliard dégénérée, Modèle de Keller-Segel

Keywords: Living tissues models, Numerical analysis, Degenerate Cahn-Hilliard equation, Keller-Segel model

Cette thèse a été préparée au

Laboratoire Jacques-Louis Lions
Sorbonne Université
Campus Pierre et Marie Curie
4 place Jussieu
75005 Paris
France

☎ +33 1 44 27 42 98
Site https://ljll.math.upmc.fr/

Je dédie cette thèse ...

À mes parents.

Le destin bat les cartes, mais c'est nous qui jouons.

Bernard Moitessier

Modèles de tissus vivants, simulations numériques et immunothérapie des cancers

Résumé

Nous étudions deux types de modèles couramment utilisés pour la représentation en temps et en espace des tumeurs : l'équation de Cahn-Hilliard pour les tissus vivants et le modèle de Keller-Segel. Les méthodes numériques que nous développons cherchent à représenter de manière précise et efficace ces équations tout en préservant leurs propriétés. Pour l'équation de Cahn-Hilliard, notre étude s'appuie sur une méthode de relaxation dont nous prouvons la convergence vers le modèle initial. Même si elles représentent mathématiquement des phénomènes physiques proches de ceux étudiés en dynamique des fluides, les équations utilisées pour les tissus vivants sont souvent différentes pour rendre compte du caractère actif des cellules. Les équations résultantes contiennent de nombreuses singularités et dégénérescences qui sont difficiles à analyser théoriquement et simuler numériquement de manière efficace. La méthode de relaxation a été introduite pour faciliter l'implémentation de nos schémas numériques ; nous proposons ainsi des schémas numériques éléments finis simples à adapter dans les codes pré-existants. Afin de préserver les propriétés des équations continues lors des simulations numériques, nous proposons des schémas numériques basés sur la Méthode de Variable Auxiliaire. L'adaptation de cette méthode pour les équations des tissus vivants n'ayant pas été réalisée, nous proposons dans cette thèse d'y remédier et d'étudier les propriétés analytiques de ces schémas numériques. Sur la base de ces travaux numériques, nous présentons l'étude de deux phénomènes biologiques. En collaboration avec des biologistes de l'Université de Nantes, nous étudions la compactification des sphéroïdes de glioblastome in-vitro en réponse à un médicament utilisé en chimiothérapie. Notre deuxième application s'intéresse à l'étude des effets physiques jouant un rôle dans l'émergence d'instabilités à la surface de certaines tumeurs invasives.

Mots clés : Modèles de tissus vivants, Analyse numérique, Équation de Cahn-Hilliard dégénérée, Modèle de Keller-Segel

Abstract

We study two classes of mathematical models currently used for the modeling in time and space of tumors: the Cahn-Hilliard equation for living tissues and the Keller-Segel model. The numerical methods we propose aim to represent these equations efficiently and accurately while preserving their properties. For the Cahn-Hilliard equation, our study is based on a relaxation method for which we prove the convergence to the original model. Even though the physical effects modeled by these equations are close to the ones studied in fluid dynamics, the equations used to model living tissues are different in order to represent the active behavior of cells. The resulting equations contain numerous singularities and degeneracies, which result in technical difficulties to analyze and simulate them efficiently. Our relaxation method has been introduced to facilitate the implementation of our numerical schemes. Hence, we propose numerical schemes that are easy to implement in already existing finite element software. In order to preserve the properties of the equations during numerical simulations, we design numerical schemes based on the Scalar Auxiliary Variable method. However, since this method has never been used in the context of models of living tissues, we study the analytical properties of our schemes. Based on these numerical works, we present two studies of biological phenomena. In collaboration with biologists from the Université de Nantes, we study the shrinking of in-vitro tumor aggregates of glioblastoma due to a certain chemotherapeutic drug. Our second study focuses on understanding the physical effects that play a role in the emergence of instabilities at the borders of certain invasive tumors. Therefore, this work aims at providing mathematical tools to biologists that give insights into underlying biological phenomena based on the Physics of cells and living matter.

Keywords: Living tissues models, Numerical analysis, Degenerate Cahn-Hilliard equation, Keller-Segel model

Laboratoire Jacques-Louis Lions
Sorbonne Université – Campus Pierre et Marie Curie – 4 place Jussieu – 75005 Paris – France

Remerciements

Cette section est une tentative de rendre hommage à toutes les personnes qui ont contribué de près ou de loin au bon déroulement de ma thèse. Certains doctorants comparent ces trois années de recherche et d'apprentissage à un voyage avec un grand nombre de difficultés à surmonter. Bien que certains moments aient été plus difficiles que d'autres, mon voyage s'est dans l'ensemble bien passé et cela est dû aux personnes avec qui ou au contact de qui j'ai eu la chance de travailler.

Tout d'abord, mes plus sincères remerciements s'adressent à mes deux directeurs de thèse. La passion communicative qui anime Benoît Perthame pour les mathématiques appliquées à la biologie ainsi que son intuition dans la recherche m'ont autant impressionné qu'elles sont devenues des sources d'inspiration pour moi. La capacité que possède Tommaso Lorenzi à comprendre et modéliser mathématiquement les phénomènes biologiques m'a profondément marqué et impressionné. J'ai été encadré par deux professeurs qui m'ont présenté les différentes facettes de la recherche en mathématiques appliquées à la biologie : entre modélisations, études théoriques, simulations numériques et comparaisons aux données biologiques, je pense que je ne pouvais espérer mieux. Ils ont aussi fait preuve à mon égard d'une patience et d'une pédagogie pour lesquelles je leur suis infiniment reconnaissant. Apprendre à leurs côtés est un réel plaisir et leur encadrement constitue le meilleur environnement possible pour le bon déroulement d'une thèse.

Je remercie aussi Alain Miranville et Christian Schmeiser pour avoir accepté de rapporter ma thèse et pour leur relecture attentive de ce manuscrit. Je suis très honoré que Marie Doumic, Katharina Schratz, Florence Hubert et Pasquale Ciarletta, que leur réputation en mathématiques appliquées aux sciences du vivant et en analyse numérique précède, participent à mon jury.

Ces trois années ont été marquées par la rencontre de nombreuses personnes avec qui j'ai eu le plaisir de travailler ou simplement échanger. Mes plus chaleureux remerciements s'adressent donc à Luís Almeida pour nos nombreuses discussions et ses précieux conseils. Ils s'adressent aussi à Katharina Schratz dont la gentillesse et la pédagogie sont complétées par des connaissances exceptionnelles en analyse numérique. De même, ils vont à Bertrand Thierry dont la sympathie n'est plus à prouver et dont les connaissances en informatique ne cesseront de m'éblouir. Mes remerciement s'adressent aussi à Jean Clairambault pour ses nombreux conseils tant sur des aspects scientifiques que pour l'organisation de conférences. Je souhaite remercier également Pasquale Ciarletta qui m'a initié sur un coin de table de la salle de conférences des Treilles à la beauté et la complexité de l'équation de Cahn-Hilliard. Équation qui m'a permis de rencontrer certains professeurs et maîtres(ses) de conférence du LMA à l'université de Poitiers : Alain Miranville, Julien Dambrine, Morgan Pierre et Cécile Taing et avec qui j'ai pu partager un bon moment à Poitiers juste avant le premier confinement. Je souhaite remercier Harald Garcke, Abramo Agosti et Maurizio Grasselli avec qui j'ai eu la chance de discuter lors de leur passage au LJLL.

J'aimerais adresser ces remerciements à Salima, Malika et Catherine pour leur aide bienveillante lors des démarches administratives, de l'organisation des missions ainsi que lors de l'organisation de la conférence qui finalement a eu lieu en visio. Je souhaite aussi remercier Kha-

shayar et Hugo pour leur aide précieuse sur tout ce qui a concerné l'informatique. Et un grand merci à Antoine pour son aide concernant les serveurs de calculs et sans qui de nombreuses simulations numériques n'auraient pas été possibles.

Le voyage en thèse ne s'effectue jamais seul et j'en viens donc à remercier ceux qui ont rendu l'expérience encore plus plaisante : Idriss, Gissell, Sophie, Elise, Jules, Lise, Emilio... et beaucoup d'autres que j'oublie certainement de mentionner, excusez-moi. Je souhaite remercier chaleureusement mes soeurs de thèse : Giorgia, Noemi... Entre jets de banane (pardon ouverture), français plus que douteux et désir de bord de meeeeeeeer, bref pas le temps de s'ennuyer. Il y aussi mes colocataires de l'ancien bureau 315 : Katia, Julia, Nicolas, Medhi et Luidi. Et les nouveaux colocataires post-déménagement : Pierre, Lucas et Charles (elle est bien la chaise hein). Sans oublier Jean-François et Pierre avec qui partager ces séances de PPG autour de Serge fut un plaisir (si on exclut les courbatures bien sûr). Il y aussi Hugo, Angélique et Emma avec qui j'ai eu le plaisir d'organiser une conférence (et tout s'est passé comme sur des roulettes bien sûr). Mes plus sincères amitiés sont aussi adressées à mon amie Zineb qui, en plus de travailler avec un acharnement dont je suis admiratif, est l'une des personnes les plus courageuses que je connaisse.

Ces trois années ont aussi été marquées par l'enseignement. Je remercie Laurent Boudin pour m'avoir aidé à dispenser des TD et TP à Polytech Sorbonne. J'ai eu la chance en faisant cela de rencontrer des professeurs qui m'ont communiqué leur passion de l'enseignement et leur méthode de travail. Je souhaite donc remercier Frédéric Paugam, Laurent Lazzarini, Frédérique Charles et Cécile Braunstein qui m'ont beaucoup appris sur l'enseignement des mathématiques et de l'informatique aux élèves ingénieurs. Bien sûr ces remerciements seraient incomplets sans mentionner les élèves de EISE, MAIN, EI-2I et ST de Polytech Sorbonne avec qui donner des TD et TP a toujours été un plaisir. Aussi, à propos de Polytech Sorbonne, je souhaite remercier Myriam Comte, qui m'a toujours soutenu durant mes études (même pendant les moments difficiles) et sans qui je n'aurais certainement pas été capable d'arriver jusque la thèse.

Cependant la thèse, comme toute période de la vie professionnelle, n'est agréable que si la vie privée l'est aussi. C'est pourquoi je souhaite remercier les personnes avec qui il est toujours plaisant d'échanger autour d'un verre (que ça soit en terrasse, au bord d'une plage ou dans un parc), ou juste décompresser et se promener. Du coup, pêle-mêle, merci à Lucile, Driss, Louis, Fadwa, Manon, Lucas, Agnès, David, William et finalement Nicolas (qui doit tout me réapprendre pour danser le rock... désolé).

Bien sûr ce travail n'aurait pas été possible sans l'amour et le soutien inconditionnel de ma famille durant ces 8 années d'études. Pour tout cela je souhaite remercier mes parents, ainsi que Pierre-Yves, Yann et Constance (tu es la prochaine sur la liste... tu verras trois ans de thèse ça passe trop vite !).

Je souhaite conclure cette section en remerciant du fond de mon coeur ma soeur de thèse Federica Bubba décédée tragiquement le 25 juin 2020. Elle était une chercheuse exceptionnelle dotée de connaissances, d'une curiosité et d'une patience qui me sont encore aujourd'hui un modèle à atteindre. Je me souviendrai toujours des nombreux conseils qu'elle m'a prodigués lors de ma première année. Mais au-delà de ça, Federica était une personne dotée d'une gentillesse remarquable avec un coeur immense. J'ai adoré effectuer les voyages à Édimbourg et Samos avec elle. Elle comme moi préférait les voyages par train et bateau plutôt que par avion... moins rapide mais souvent le paysage est plus beau au sol et lorsqu'on va lentement.

Federica, spero dal profondo del mio cuore che dove sei oggi, tu sia felice e che riposi in pace. Ti ringrazio ancora per tutto.

Table des matières

Chapter 1

Introduction

1.1 Motivations

Mechanobiology is the science that focuses on the mechanics of biological systems. This is a science at the interface of Biology, Mechanics, and Mathematics. This manuscript focuses on the design, analysis and numerical simulation of mathematical models of living tissues. Thereby, this work belongs in the active research field of the mathematics of Mechanobiology. These models describe the evolution of cells in time and space. In the present study, we focus on the representation of living matter as a continuum, adopting a macroscopic point of view. The structures depicted are referred to as tissues but under this definition falls many cells' arrangements as long as they are observable with the naked eye. These models do not aim to give a precise description of the microbiology of cells but are focusing on average quantities and phenomena occurring at the scale of the structure. Therefore, the tissue mechanics that affect its shape, organization, and growth are the main focus of these models. Phenomena that can be represented are the collective movement of cells, the growth of the tissue, the formation of patterns, etc. The main objective of researches in mathematical Mechanobiology is to propose to biologists mathematical tools that give insights into underlying biological phenomena based on the Physics of cells and living matter in general. The models of living tissues are also used to study pathologies such as inflammations and cancer. Indeed, the previous definition of living tissue encompasses tumors. A tumor is a collective organization of cells, and the mechanical effects exerted by and on it are of primary importance. Furthermore, the effect of treatment on tumors can also be included in the model. A second objective of the work in mathematical Mechanobiology is to provide medical teams with tools for predicting the behavior of a specific tissue (healthy or tumorous) under certain constraints such as mechanical stresses or drugs.

Due to cells' active behavior, mathematical models of living tissues often pose technical difficulties both for analytical and numerical works. They are often of nonlinear type and sometimes exhibit a hyperbolic behavior, i.e., forming discontinuous interfaces even if the initial solution is smooth. Other effects such as backward diffusion or degeneracy of a mobility coefficient lead to additional analytical difficulties. On the numerical side, the main issue is to keep the properties of the solution of the continuous model during numerical simulations. Indeed, the mathematical model is an approximated representation of the biological tissue, and the numerical method is another approximation level. To make reliable predictions from the numerical simulations, it is necessary to design a numerical scheme that is the best approximation possible of the continuous model. Therefore, the design of the numerical scheme must be oriented to preserve the essential properties of models of living tissues that are: the monotonic decay of the associated energy,

the preservation of the positivity of the densities, the conservation of the initial mass of cells if proliferation does not occur, etc.

However, accuracy must not be the only focus. To apply a mathematical model to a concrete study in relation with Biology or Medicine, it is necessary to test many scenarios and, hence, run many numerical simulations. To do so, the numerical method must be efficient, and its computational cost must be reduced to its minimum for each simulation. The tradeoff question between accuracy and efficiency is at the center of every numerical work with the aim of concrete applications.

Based on these observations, in this manuscript, we focus on the design of accurate and efficient novel numerical methods for some nonlinear partial differential equations used for the mathematical representation of living tissues. Mainly, we work with models used to represent in vivo and in vitro tumors that are a pathology of living tissue.

1.2 Mathematical representation of living tissues

1.2.1 First steps of mathematical modelling

We present the formulation of a prototypal model of $N \in \mathbb{N}^*$ components from which we can derive models that include precise effects. We define the domain $\Omega(t) \subset \mathbb{R}^d$ as the tissue or a part of it, and where $d = 1, 2, 3$ is the dimension and t is time. The boundary of $\Omega(t)$ is denoted by $\partial\Omega(t)$ and is assumed to be sufficiently regular. In the following, we adopt the Eulerian point of view: for a fixed point in space, we observe what happens without moving x. The second point of view (that we will not consider in this manuscript) is the Lagrangian or material point of view, where the observer follows the flow i.e. the material point $X \in \Omega(t)$ of the domain moves with respect to the flow given by a certain velocity. To describe how the different forces act on the tissue, we take an arbitrary volume $V(t) \subset \Omega(t)$. We define $B_r(x)$ as the ball of center x and of radius $r \geq 0$. We define the mass density of the i-th component at $(t, x) \in R^+ \times \Omega$ in Eulerian coordinates by

$$\rho_i(t, x) = \lim_{r \to 0} \frac{M_i(B_r(x))}{V(B_r(x))},$$

where $M_i(B_r(x))$ is the mass of the i-th component inside the ball $B_r(x)$ and $V(B_r(x))$ is the volume of the latter. We also define relative quantities such as the relative mass densities

$$\tilde{\rho}_i(t, x) = \lim_{r \to 0} \frac{M_i(B_r(x))}{V_i(B_r(x))},$$

where $V_i(B_r(x))$ is the volume occupied by the i-th component inside the ball $B_r(x)$, and the volume fraction of the i-th component

$$n_i(t, x) = \lim_{r \to 0} \frac{V_i(B_r(x))}{V(B_r(x))}.$$

We also consider the mass fraction c_i defined by

$$c_i(t, x) = \lim_{r \to 0} \frac{M_i(B_r(x))}{M(B_r(x))}.$$

Therefore, it is easy to verify the relation

$$\rho c_i = \tilde{\rho}_i n_i, \quad \text{for} \quad i = 1, \ldots, N.$$

The starting point of the mathematical models that we study is the conservation equation in integral form

$$\int_{V(t)} \frac{\partial \rho_i(t,x)}{\partial t} + \nabla \cdot \mathbf{J}_i(t,x)\ \mathrm{d}x = \int_{V(t)} S_i(t,x)\ \mathrm{d}x, \tag{1.1}$$

where $\mathbf{J}_i(t,x) = \mathbf{J}_i(\rho, \nabla\rho, \mathbf{v})$ is the net flow of the components across the boundary of the arbitrary volume $V(t)$. This flux is defined as a function of the density of the i-th component, of its gradient and of a velocity field $\mathbf{v}_i(t,x)$ which needs to be defined using a constitutive relation or given by an equation. However, to be able to represent the Physics of the tissue, one needs to use laws of mechanics and thermodynamics to derive a set of physically relevant equations and explain the different constitutive relations. In (1.1), the term $S_i(t,x)$ represents the growth of the tissue but also its degradation. This proliferation effect depends on many aspects occurring both at the scale of the cells and the tissue, such as the stress in the tissue, the amount of available nutrients, or the available space in the neighborhood of dividing cells.

The integral form of the continuity equation is useful to give a description of the different forces acting on the arbitrary volume $V(t)$. However, since the equation (1.1) is satisfied for any volume $V(t)$, we use the local form of this equation

$$\frac{\partial \rho_i(t,x)}{\partial t} + \nabla \cdot \mathbf{J}_i(t,x) = S_i(t,x), \tag{1.2}$$

In the following, we present the derivation of two essential models of living tissues—the Cahn-Hilliard model for Biology and the Keller-Segel model. The first one is used to represent tissues as a multiphase fluid and describes the attractive and repulsive interactions of cells. As a result, the model can reproduce the formation of patterns and is used to represent tumors as fluids. The second is modeling chemotaxis: a type of movement that cells exhibit in nature. Indeed, cells have the capacity to sense their micro-environment, and their migration is driven by signals. The Keller-Segel model represents cells' movement toward zones of a large concentration of certain chemicals called chemoattractants and due to random Brownian motion.

1.2.2 The Cahn-Hilliard model for modelling of tissues and tumours

Derivation. A detailed description of the derivation of the Cahn-Hilliard equation (CH in short) from the different mechanics and thermodynamics laws requires lengthy calculations. For the sake of simplicity, we give here a simple description of the Cahn-Hilliard equation for the mixture of two incompressible fluids. We also assume that the tissue under investigation is not growing nor degrading. This model is relatively simple and focuses only on the organization of two types of cells. However, as we will see in the following of this manuscript, it already gives a good description of tissues (especially tumors) and is at the center of many research pieces.

To satisfy the previous assumptions, we set $N = 2$ and $S_i(t,x) = 0$ (for $i = 1, 2$). For the representation of tissues, we can assume that cells inside the phase $i = 1$ are cells constituting the tissue of interest (or the tumor), and the phase $i = 2$ is used to represent the rest of the cells of the micro-environment. From the incompressibility assumption, the continuity equation (1.2) can be simplified and written to follow the evolution of the mass fraction c_i or the volume fraction n_i. We present the case where we focus on volume fractions, and since these quantities satisfy

$$n_1 + n_2 = 1,$$

we define the order parameter $n = n_1$. Hence, in order to obtain the evolution of both volume

fractions, it is now only necessary to solve the continuity equation

$$\frac{\partial n}{\partial t} + \operatorname{div}(\mathbf{h}) = 0, \tag{1.3}$$

where we have used the fact that $\mathbf{J}$ in (1.2) was equal to $\mathbf{J} = \rho_i \mathbf{h}_i$ along with the fact that the fluids are incompressible, i.e. $\operatorname{div}(\rho_i) = 0$ for all $i = 1, 2$. Before giving the constitutive relation for the flux $\mathbf{h}$, we first describe the energy associated to this model and use some of basic thermodynamics quantities.

Indeed, the formulation of the free energy associated to the system is often considered the starting point to give a simple derivation of the Cahn-Hilliard model. As in its original description, the Ginzburg-Landau free energy is given by

$$\mathcal{E}[n](t) := \int_\Omega \left(\frac{\gamma}{2} |\nabla n|^2 + \psi(n) \right) \mathrm{d}x.$$

The free energy density is the sum of two important terms. The surface tension $\frac{\gamma}{2} |\nabla n|^2$ is a force occurring at the interface between the two phases. This terms has the effect to penalize large gradient of the order parameter and tends to make the interface between the two phases smooth. Hence, the length of this diffuse interface is given by $\sqrt{\gamma}$. The second term $\psi(n)$ is related to the mechanical interactions between the cells, and is called the homogeneous free energy. Attractive and repulsive forces for the two cell types are represented by this term. Therefore, attraction occurs when $\psi''(n) < 0$ and repulsion when $\psi''(n) > 0$. For volume fractions such that $\psi''(n) = 0$, we say that the mixture is at equilibrium, i.e. attractive and repulsive forces balance out.

We define $\mu = \mu(n, \nabla n)$ the chemical potential as the variational derivative of the free energy with respect to the order parameter

$$\mu = \frac{\delta \mathcal{E}}{\delta n}$$

Going back to the definition of the net flux $\mathbf{h}$, and using a generalized Fick's law, we obtain

$$\mathbf{h} = -b(n) \nabla \mu,$$

where $b(n)$ is a mobility coefficient to represent the active movement of cells. Altogether, and assuming zero-flux boundary conditions, the original Cahn-Hilliard equation for diphasic fluids reads

$$\begin{cases} \frac{\partial n}{\partial t} = \operatorname{div}\left(b(n) \nabla \left(-\gamma \Delta n + \psi'(n) \right) \right), & t > 0,\ x \in \Omega, \\ \frac{\partial \mu}{\partial \nu} = \frac{\partial n}{\partial \nu} = 0, & t > 0,\ x \in \partial\Omega, \end{cases} \tag{1.4}$$

where ν is the outward normal vector to the boundary $\partial\Omega$.

The Cahn-Hilliard for Biology. This equation found its original application in the context of material sciences. Cahn and Hilliard [48, 47] first proposed the equation to represent the separation of phases occurring in binary alloys during a sudden cooling, assuming isotropy and constant temperature. Indeed, the equation can model the different stages of phase separation: from the formation of microscopic structures (i.e. spinodal decomposition) to the coarsening of them to form large arrangements. Later, the Cahn-Hilliard equation has been used to represent many phenomena in Physics, such as dealloying occurring in corrosion [81], thin films [191], image processing [53], or even about the formation of the rings of Saturn [196]. The previous references do not cover the extensive literature that exists for each of these subjects. However, the interested reader can find more references in the review book of Miranville [147].

In this manuscript, we are interested in its application to Biology, especially to represent tissues and tumors. Due to qualitative similarities between the formation of patterns in nature and the processes of phase separation occurring in materials, researchers started to use the Cahn-Hilliard equation as a phenomenological model for biological applications, such as population dynamics [62], wound healing [126], tumor growth [203, 90] or even the organization of mussel banks [138]. In the previous references, a source term $g(n)$ is often considered inside the equation to represent growth of the tissue or its degradation, leading to the generalized Cahn-Hilliard equation [59]

$$\frac{\partial n}{\partial t} = \operatorname{div}\left(b(n)\nabla\left(-\gamma\Delta n + \psi'(n)\right)\right) + g(n), \quad t > 0,\, x \in \Omega. \tag{1.5}$$

Both [62] and [126] used the Cahn-Hilliard to produce results close to observed phenomenon without focusing on the consistency of the model with their system thermodynamics. The first work that proposed a mechanically and thermodynamically consistent derivation of the CH equation for Biology is [203]. The authors derived a system of CH-type equations using thermodynamics laws to derive constraints for the constitutive relations of their model. Their model is capable of representing multi-species systems, and they considered as a test case the representation of a system with four species: viable tumor cells, dead tumor cells (necrotic core), healthy cells, and the rest of the micro-environment. In the previous references that propose a CH model for living tissues, the mobility coefficient (i.e. $b(n)$ in (1.5)) is taken to be a constant and the potential $\psi(n)$ is double-welled and logarithmic, i.e. it describes the segregation of the two different types of cells. For this application, a thermodynamically consistent potential used in many research pieces is

$$\psi(n) = \frac{1}{2}n\ln n + (1-n)\ln(1-n) - (n-\frac{1}{2})^2,$$

but is often approximated by a polynomial function for simplicity.

Later, other forms for the mobility and potential have been proposed to give a better representation of the mechanics of cells inside tissues. Considering a constant mobility in the CH model does not seem biologically relevant. Indeed, to represent the clustering of cells, a degenerate mobility is more relevant as pointed in [8, 6], and consists in taking the mobility to be zero in pure phase i.e. in zones where cells are too overcrowded ($n = 1$ and $n = 0$). A possible example is

$$b(n) = n(1-n)^2.$$

Featuring this kind of mobility the CH is often referred to as *the degenerate Cahn-Hilliard model* (DCH in short). Another modification that renders the model biologically relevant for certain applications, concerns the potential $\psi(n)$. As said previously, this potential is often taken to be a double-well logarithmic function with singularity points at the pure phases. However, as proposed in [46], a single-well logarithmic potential seems to be more relevant for the modeling of cancer. Indeed, tumor cells have the capacity to spontaneously form aggregates, and are often the only active cells in the experiments. For example, modelling the formation of in-vitro spheroids of cancerous cells, one can consider the two phases of the fluid to be the tumor cells and the other phase is the inactive gel that serves as the culture medium. For this kind of study, a phenomenological potential is

$$\psi(n) = -(1-n^*)\ln(1-n) - \frac{n^3}{3} - (1-n^*)\frac{n^2}{2} - (1-n^*)n,$$

in which $0 < n^\star \leq 1$ is a parameter used to represent the maximal density at which aggregates are stable, i.e. attraction and repulsion forces balance out.

Preview. In Part I, we work with the DCH model with a single-well logarithmic potential. Compared with the CH model original form (with constant mobility and a double-well logarithmic potential), these modifications induce numerous difficulties on both the analytical and numerous sides. To tackle the resulting issues, we propose a relaxation of the model and design a structure-preserving numerical method to simulate it.

The DCH with a single-well logarithmic homogeneous free energy has been used for the modelling of skin cancer [55, 56, 61], and of glioblastoma [9]. In both of these studies, the results of numerical simulations of the model were in good agreement with biological observations and experimental results. Furthermore, in [10], the authors used their mathematical and algorithm framework, which involves the DCH model in a concrete study with data coming from a patient suffering from glioblastoma. The numerical simulations were able to represent the evolution of the tumor of the patient even under treatment. The mathematical model was able to match the volume and boundaries of the tumor observed in images realized by Magnetic Resonance Imaging (MRI).

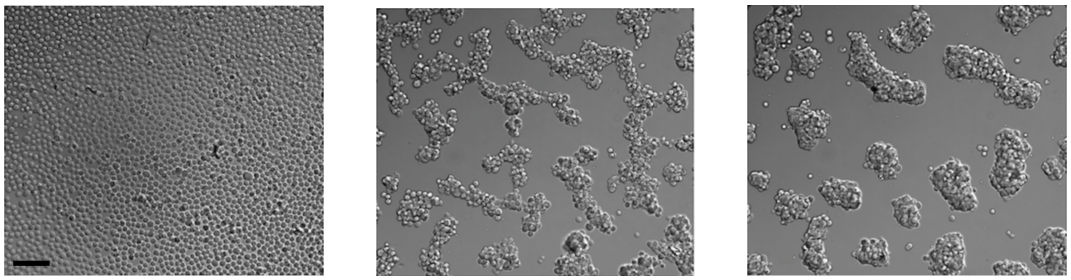
Although the previous Cahn-Hilliard framework provides a good representation of a tumor in a healthy tissue or in vitro, as a diphasic fluid, it is necessary to consider a larger number of cell types and include particular mechanical effects in some applications. To this end, in [98] the authors derived a system of equations that comprises a CH-type equation to represent healthy and tumor cells, coupled to an equation for the diffusion of the nutrients in the tissue. This model is derived based on simple thermodynamics principles and represents nutrient diffusion, chemotaxis, active transport, adhesion between cells, apoptosis (i.e., death of the cells), and proliferation. Therefore, in the equation for the chemical potential, the effect of nutrients is taken into account to represent the chemotactic movement. A velocity field is also taken into account to represent non-active movement (i.e., advection). It is given by Darcy's law which states that the movement is towards zones of lower pressure. The extension of the model for multispecies systems has been proposed in [97]. Numerical simulations of this model give a good qualitative representation of biological phenomena, such as the emergence of a necrotic core inside the tumor. In the two previous models, the transport due to the fluid movement was given by Darcy's law. However, to account for the fluid viscosity, a Cahn-Hilliard-Brinkman model has been proposed for tumor growth [72].

Preview. In the CH-type models discussed so far, tissues are considered components of an incompressible fluid. This assumption leads to a good representation of Biology, as we have seen. However, in some cases, it could be interesting to keep compressible effects. In Chapter 5, we derive a model for a compressible diphasic fluid to represent two populations of cells or two different tissues. The compressibility of the cells is maintained, and the complete system that we derive is a Cahn-Hilliard equation coupled with a compressible Navier-Stokes equation for the velocity field

1.2.3 The Keller-Segel model and the volume-filling approach

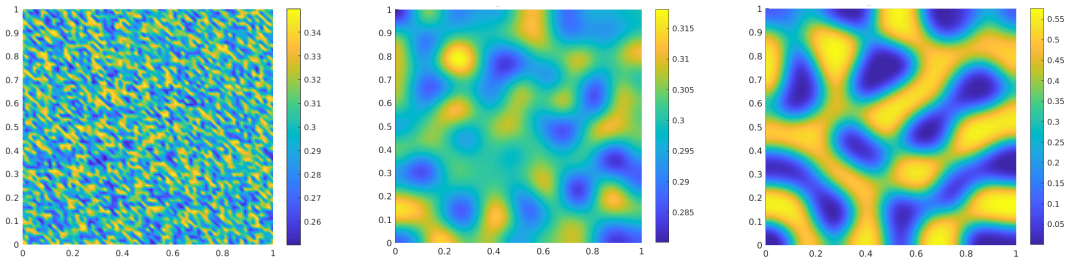
Derivation. As for the original proposition of the Cahn-Hilliard equation, the Keller-Segel model (KS in short) has first been introduced as a phenomenological model to describe cells' movement toward a specific signal called the chemoattractant. This type of motion for cells and

Qualitative comparison between biological experiments and numerical simulation of the degenerate Cahn-Hilliard model.
Since the same processes occur during phase separation in materials and pattern formation for self-organization of cells, the CH model has been considered as a good model for biological applications. As an example, we present the results of experiments for the formation of spheroids of tumor cells and compared them qualitatively with numerical experiments of the CH model (the numerical simulations have been performed with the scheme proposed in [166]). Figure 1.1 is taken from [5] and shows the formation of structures for glioma cells of rats. From a random distribution (Figures 1.1a and 1.1b), cells tend to form rapidly small aggregates. The second phase consists of the merging of some of these small aggregates to form larger structures (Figure 1.1c).



(a) T = 0 (b) T = 6h (c) T = 12h

Figure 1.1 – Aggregation of glioma cells from rat during time reproduced from [5] (CC BY-NC 3.0).



(a) $t = 0$ (b) $t = 0.5$ (c) $t = 1$

Figure 1.2 – Simulation of the Cahn-Hilliard equation from the numerical method proposed in [167]
Figure 1.2 depicts three states during a numerical simulation of the DCH model with a single-well logarithmic potential. This model has been used in its non-dimensionalized form and, hence, time is rescaled and $t \in [0, 1]$. Figure 1.2a shows the initial condition, a uniform distribution of the order parameter n (i.e., relative cell density or volume fraction) around the value $n_0 = 0.3$. Figure 1.2b is a plot of the order parameter at $t = 0.5$, and we see that small structures are formed already. At the end of the simulation (Figure 1.2c), these small aggregates have merged to form larger structures.
Comparing the experiments in Figure 1.1 with the numerical simulations on Figure 1.2, we clearly observe a qualitative match in terms of behavior of cells and structures formed.

bacteria is called chemotaxis. Starting from the prototypal model (1.2), we define the flux for each population by

$$\mathbf{J}_i = -D_i(\rho_i)\nabla\rho_i + \chi(\rho_i)\mathbf{v}_i, \tag{1.6}$$

where $D_i(\rho_i)$ is a density-dependent diffusion coefficient, $\chi(\rho_i)$ a function which gives the strength of the chemotactic movement. Therefore, chemotaxis is described mathematically by a specific velocity field $\mathbf{v}_i$. Hence, $\mathbf{v}_i$ is a function of one or many other component $j \neq i$. For a single cell population, if we consider a flux of the form (1.6) and the chemoattractant is given by a known function $F(x)$, the model (1.2) reduces to the Fokker-Planck equation. Generally, the concentration of chemoattratant is given by a diffusion equation. To simplify the derivation, we consider in the following the biological situation of a unique cell population $i = 1$, and a single chemoattractant $i = 2$. The movement of the cells is a combination of random Brownian motion and chemotaxis. We designate by $u = \rho_1$ the density of cells and by $c = \rho_2$ the concentration of the chemoattractant that diffuses in the domain. The two fluxes (1.6) are given by

$$\begin{aligned} J_1 &= -D_1(u,c)\nabla u + \chi(u,c)\nabla c, \\ J_2 &= -D_2(u,c)\nabla c. \end{aligned}$$

Therefore, the general Keller-Segel model reads

$$\begin{cases} \frac{\partial u}{\partial t} &= \operatorname{div}\left(D_1(u,c)\nabla u\right) - \operatorname{div}\left(\chi(u,c)\nabla c\right) + S_1(u,c), \\ \frac{\partial c}{\partial t} &= \operatorname{div}\left(D_2(u,c)\nabla c\right) + S_2(u,c), \end{cases} \tag{1.7}$$

and is often supplemented by zero-flux boundary conditions

$$\frac{\partial u}{\partial \nu} = \frac{\partial c}{\partial \nu} = 0. \tag{1.8}$$

Function $S_1(u,c)$ describes the growth of the proliferating cell population that depends on the local cell density (since the proliferation of existing cells gives growth) and, possibly, on the chemoattractant concentration. The second source term $S_2(u,c)$ represents the production and decay of the chemoattractant. Both functions $D_1(u,c)$ and $D_2(u,c)$ are diffusion coefficients of respectively the cells and the chemoattractant. These two latter functions are often considered to be constant coefficients. Particular attention will be given in the following to the function $\chi(u,c)$ referred to as the chemosensitivity and describes the strength of chemotaxis. This function can depend both on the density of cells (as for the mobility in the CH model case that we have seen before) and on the concentration of chemoattractant. In the form (1.7), the KS model is a system of two parabolic equations and is often referred to as the *parabolic-parabolic Keller-Segel model*. If the diffusion of the chemoattractant is very fast compared to the Brownian motion of the cells i.e. $\frac{D_1(u,c)}{D_2(u,c)} \equiv 0$, then the parabolic-parabolic KS model is approximated by

$$\begin{cases} \frac{\partial u}{\partial t} &= \operatorname{div}\left(D_1(u,c)\nabla u\right) + \operatorname{div}\left(\chi(u,c)\nabla c\right) + S_1(u,c), \\ 0 &= \operatorname{div}\left(D_2(u,c)\nabla c\right) + S_2(u,c), \end{cases} \tag{1.9}$$

and is called the *parabolic-elliptic Keller-Segel model* (since the second equation is now of elliptic type). Further approximations of this system are possible, but are not covered in this manuscript.

The Keller-Segel model. In the article [164], Patlak studied the effect of an external bias on the movement of particles. From a model that describes the cells individually, Patlak obtained a modified Fokker-Planck equation. Independently, Keller and Segel, in the 70's [123, 122, 124]

worked on the modeling of the aggregation of cellular slime mold called acrasiales. Indeed, biologists observed a long-range effect in the morphogenesis of these aggregates depending on chemical cues present in the environment. Therefore, incorporating the different reactions for their application, they wrote a four equations system that they reduced to the general system (1.7). This was the first time this model was proposed in this form. Since then, mathematicians have started to analyze its analytical properties, and find new ways to simulate it efficiently.

One of the simplest version of the general model (1.7), is the *minimal Keller-Segel model*, and is obtained by taking $D_1(u,c) = D_1 \geq 0$, $D_2(u,c) = D_2 \geq 0$, and $\chi(u,c) = u$. Altogether, the model reads

$$\begin{cases} \frac{\partial u}{\partial t} &= D_1 \Delta u + \operatorname{div}(u \nabla c), \\ \tau \frac{\partial c}{\partial t} &= D_2 \Delta c + u - c, \end{cases} \tag{1.10}$$

where $\tau = \frac{D_1}{D_2}$ describes how fast the chemoattractant diffuses compared to the Brownian motion of the cells. This relatively simple model is capable to represent the aggregation of a constant mass of cells due to a chemical signal c produced by the cells themselves. The properties of the solutions of this model have been studied by many authors. One of the most important results concerns the potential blow-up of the solution u in finite time. Indeed, from [158] we know that if $d = 1$, the model (1.10) has a global weak solution, and u remains bounded. However, for dimension $d \geq 2$, it exists a critical mass M such that a global weak solution exists if

$$\int_\Omega u(0,x)\, \mathrm{d}x \leq M.$$

The precise value of this critical mass has been found for $d = 2$ [158], and is $M = 4\pi$. For $d \geq 3$ [51], if $\|u_0\|_{L^{\frac{d}{2}}(\Omega)}, \|\nabla c_0\|_{L^d(\Omega)} \leq \varepsilon$ where $\varepsilon > 0$, it exists a global and bounded weak solution $\{u, c\}$. From numerical simulations, we observe that the system forms patterns that appear in Turing-type instabilities. Mainly, from a uniform initial cell density, spikes of large cell density will form, and if the simulation is run long enough, all the cells aggregates in a single sharp peak. However, it is easy to understand that this behavior is not biologically relevant. Indeed, the formation of overcrowded zones is a scenario that cells tend to avoid. To solve this issue and represent chemotaxis in living organisms, variants of the KS model (1.7) have been proposed.

To avoid the blow-up of the solution, particular forms of the functions $D_1(u,c)$, $D_2(u,c)$, and $\chi(u,c)$ can be chosen. Indeed, in [162], Painter and Hillen proposed modifying the functions for chemosensitivity to take into account the effects of "volume-filling" and "quorum-sensing". The first effect is motivated by the fact that cells have a finite size. Hence, they cannot aggregate indefinitely at a certain point, i.e. the space available for new cells to move at a specific location decreases as the density increases. Therefore, taking into account this assumption, cells tend to form aggregates that have a finite saturation value. The second effect, which is "quorum-sensing", captures how cells behave to achieve homeostasis. Indeed, biological tissues organize themselves in such a way as to avoid excessive cell densities that may result in a depletion of important nutrients and, hence, necrosis. To model this effect, a supplementary chemical w is introduced in the system to allow the cells to sense if the zones are overcrowded and change, consequently the chemosensitivity. In [162], the authors started from a model with continuous-time and discrete space. The domain is divided into discrete locations, and cells have a probability to jump to a neighboring location depending on the chemoattractant. The authors added a set of rules to take into account the effect of volume-filling and quorum-sensing. Then, they formally derived a PDE model and retrieved a Keller-Segel system with a particular form for the chemosensitivity.

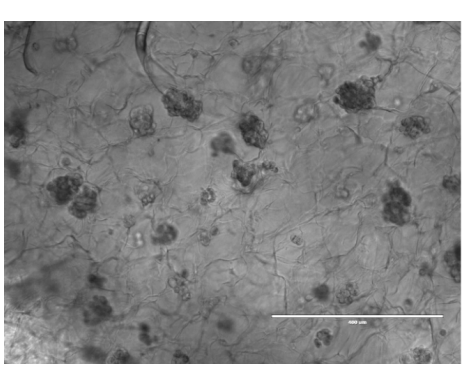
Indeed, a simple modification of the original system (1.10) can prevent the blow of the solution. Taking $\chi(u) = \chi_c u(1 - u/\overline{u})$ (where χ_c is a positive constant and $\overline{u}$ is the saturation

density at which aggregates are stable) instead of $\chi(u) = u$, takes into account the effect of volume-filling. This model reads

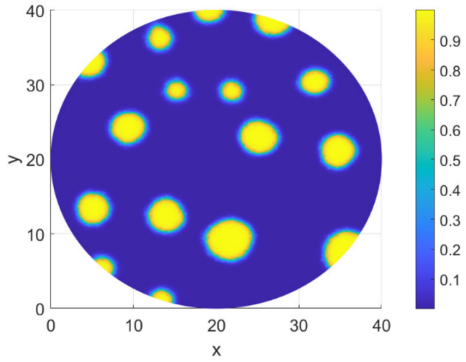
$$\begin{cases} \frac{\partial u}{\partial t} & = D_u \Delta u + \chi_c \text{div}\left(u(1-u)\nabla c\right), \\ \tau \frac{\partial c}{\partial t} & = D_c \Delta c + u - c. \end{cases} \tag{1.11}$$

Preview. In Chapter 6, we work with the volume-filling parabolic-parabolic KS model (1.11). This nonlinear system represents the aggregation of cells with a saturation value $\overline{u} = 1$. This system induces difficulties at the discrete level, and we propose in Chapter 6 a novel numerical method.

Qualitative comparison between biological experiments and numerical simulation of the volume-filling Keller-Segel model.
As explored in [45], chemotaxis could be an explanation of the self-organization of tumor cells and to the formation of in-vitro spheroids. As an example, we redo the numerical experiments of [45] using our code from [13]. The biological experiments are taken from [45] as well. Figure 1.3a shows the spheroids of tumor cells from breast cancer (MCF-7) in hydrogel. Figure 1.3b is the end state of a simulation of the KS model with volume-filling, and we see round zones where the cell density is large and between these aggregates zones where no cells are present. We can conclude of a qualitative good agreement between the experiments and the numerical simulations.



(a)



(b)

Figure 1.3 – Spheroids of MCF-7 taken from [45] (left, with permission from the journal to reproduce the figure), and numerical simulation of the Keller-Segel model with volume-filling (right)

Analytical properties of the Keller-Selel model with volume-filling. Apart from the prevention of the blow-up of the solution, the KS model (1.11) has other interesting properties. This model has been analyzed by Hillen and Painter [162] in which they prove the existence of non-negative global weak solutions using the theory of semi-groups. In our case, and as for the

Cahn-Hilliard model, we are interested into the gradient-flow structure of the KS model. Indeed, the energy associated to the model and its dissipation are given by

$$\mathcal{E}[u,c](t) = \int_\Omega \frac{D_u}{\chi_c}\left[u\log u - (u-1)\log(1-u)\right] - uc + \frac{1}{2}\left(|\nabla c|^2 + \alpha c^2\right) + C \,\mathrm{d}x,$$
$$\frac{\mathrm{d}\mathcal{E}[u,c](t)}{\mathrm{d}t} = -\int_\Omega \chi(u)\left|\frac{\delta\mathcal{E}}{\delta u}\right|^2 + \left|\frac{\delta\mathcal{E}}{\delta c}\right| \,\mathrm{d}x.$$

In its gradient-flow formulation, the model reads

$$\begin{cases} \frac{\partial u}{\partial t} & = \nabla\cdot\left(\chi_c\varphi(u)\nabla\frac{\delta\mathcal{E}}{\delta u}\right), \\ \tau\partial_t c & = -\frac{\delta\mathcal{E}}{\delta c}. \end{cases} \tag{1.12}$$

This formulation of the KS model will be useful when designing a numerical scheme that preserves its structure and analytical properties.

1.3 General assumptions and preliminaries

Assumptions on the domain Ω. To set our problems and to define the numerical schemes for these models, we first give details about the domain Ω. For dimension $d = 1$, Ω is an open, bounded interval i.e. $\Omega \subset \mathbb{R}$. For dimension $d \geq 2$, Ω is an open, bounded and connected set of $\mathbb{R}^d$. Generally, in our biological applications, we take $d = 1, 2, 3$. For $d = 2, 3$, we need to give some assumptions on the domain boundary $\partial\Omega$. We assume that the domain has a Lipschitz boundary (see Definition 1) or has a C^k boundary with $k \geq 1$, which is a necessary condition to use important results such as Sobolev injections or Poincaré-Wirtinger inequality.

Definition 1 (Lipschitz domain [4]) *A domain Ω is said to be a Lipschitz domain if: there are $\alpha, \beta > 0$, a finite number of coordinate systems $x^r = (x^{r\prime}, x_d^r)$, $1 \leq r \leq R$, where $x^{r\prime} \in \mathbb{R}^{d-1}$, and $x_d^r \in \mathbb{R}$. There are also R local maps ϕ^r that are Lipschitz continuous on their definition domain $\{x^{r\prime} \in \mathbb{R}^{d-1}; |x^{r\prime}| < \alpha\}$ such that*

$$\partial\Omega \bigcup_1^R \{(x^{r\prime}, x_d^r); x_d^r = \phi^r(x^{r\prime}); |x^{r\prime}| \leq \alpha\},$$

$$\{(x^{r\prime}, x_d^r); \phi^r(x^{r\prime}) < x_d^r < \phi^r(x^{r\prime}) + \beta; |x^{r\prime}| \leq \alpha\} \subset \Omega, \quad \forall r,$$
$$\{(x^{r\prime}, x_d^r); \phi^r(x^{r\prime}) - \beta < x_d^r < \phi^r(x^{r\prime}); |x^{r\prime}| < \alpha\} \quad \subset \mathbb{R}^d \setminus \overline{\Omega}, \quad \forall r,$$

where $|x^{r\prime}| \leq \alpha$ means that $|x_i^{r\prime}| \leq \alpha$ for all $1 \leq i \leq d-1$.

The above definition is a precise description of a Lipschitz boundary, that formally means that every $x \in \partial\Omega$ has a neighborhood U_x whose intersection with $\partial\Omega$ is the graph of a Lipschitz continuous function (i.e. the domain has locally a Lipschitz boundary). Definition 1 also requires that the domain is located on one side of its boundary, and therefore, we exclude domains with cracks or slits.

Therefore, in the rest of this manuscript, when the boundary of the domain is referred to as being "smooth enough", we implicitly assume that it satisfies at least Definition 1. A stronger assumption is to define the domain to be of class C^k, which means that its boundary is the graph of a C^k function.

Notations. We indicate the usual Lebesgue and Sobolev spaces by respectively $L^p(\Omega)$, $W^{m,p}(\Omega)$ with $H^m(\Omega) := W^{m,2}(\Omega)$, where $1 \leq p \leq +\infty$ and $m \in \mathbb{N}$. For a general function f, the corresponding norms are denoted by

$$\|f\|_{W^{m,p}(\Omega)} = \|f\|_{m,p,\Omega}\,, \quad \|f\|_{H^m(\Omega)} = ||f||_{m,\Omega},$$

with the semi-norms

$$\|D^m f\|_{L^p(\Omega)} = |f|_{W^{m,p}(\Omega)} = |f|_{m,p,\Omega}\,, \quad \|D^m f\|_{L^2(\Omega)} = |f|_{H^m(\Omega)} = |f|_{m,\Omega}\,,$$

where D^m denotes the m-th derivative of f.

The standard L^2 inner product will be denoted by $(\cdot,\cdot)_\Omega$ and the duality pairing between $(H^1(\Omega))'$ and $H^1(\Omega)$ by $<\cdot,\cdot>_\Omega$.

Inequalities and compactness in Banach spaces. We recall important inequalities that will be used throughout this manuscript. Here, we give their statement without giving their proof which could be out of the scope of this manuscript (details can be found in [4, 134]). We assume that the domain Ω satisfies the assumptions of Definition 1.

Proposition 2 (Poincaré-Wirtinger inequality) *Let $1 \leq p < \infty$, for a function $u \in W^{1,2}(\Omega)$, we have*

$$\|u - u_\Omega\|_{L^2(\Omega)} \leq C\,\|\nabla u\|_{L^2(\Omega)}\,,$$

where u_Ω is the average value of u on Ω i.e.

$$u_\Omega = \frac{1}{|\Omega|}\int_\Omega u(x)\,\mathrm{d}x,$$

and $|\Omega|$ is the measure of the domain.

Proposition 3 (Continuous embeddings of Sobolev spaces [4]) *For $1 \leq p, q < \infty$, and k, l being positive integers such that $k > l$ such that*

$$\frac{1}{p} - \frac{k}{d} = \frac{1}{q} - \frac{l}{d},$$

and we have $W^{k,p}(\Omega) \hookrightarrow W^{l,q}(\Omega)$.
A particular case is $k = 1$ and $l = 0$, for which we have $W^{1,p}(\Omega) \hookrightarrow L^q(\Omega)$.
Furthermore, for $\mu = 1 - \frac{d}{p}$, we have $W^{1,p}(\Omega) \hookrightarrow C^{0,\mu}(\Omega)$.

Proposition 4 (Compact embeddings of Sobolev spaces [4]) *For $1 \leq p, q < \infty$, and for j, m being positive integers, we have*

$$W^{j+m,p}(\Omega) \subset\subset W^{j,q}(\Omega), \quad \text{if} \quad 0 < n - mp, \quad \text{and} \quad j + m - \frac{n}{p} \geq j - \frac{n}{q},$$

and

$$W^{j+m,p}(\Omega) \subset\subset C^j(\overline{\Omega}), \quad \text{if} \quad mp > n.$$

Proposition 5 (Lions-Aubin Lemma [182, 134]) *Let X, Y, Z be Banach spaces with a compact embedding $X \subset\subset Y$, and a continuous embedding $Y \hookrightarrow Z$. Then, we have the compact embedding*

$$\{u \in L^2(0,T;X) | \frac{\partial u}{\partial t} \in L^2(0,T;Z)\} \subset\subset L^2(0,T;Y),$$

and

$$\{u \in L^\infty(0,T;X) | \frac{\partial u}{\partial t} \in L^2(0,T;Z)\} \subset\subset C([0,T];Y).$$

1.4 Numerical simulation: foundations

To obtain some insights on the behavior of the solution of the partial differential equations, numerical simulations can be useful. In our application to living tissues, numerical simulations can illustrate key cellular processes and allow for a better understanding of key cellular processes that underpin the dynamics of cells in living organisms and may allow for reliable qualitative and quantitative predictions to be made.

This section aims to describe the mathematical foundations of the numerical methods that we use for our applications. For each of the models described in the previous section, the unknowns are functions of time and space $(t,x) \in \Omega_T = [0,T] \times \Omega$.

We use the method of lines to discretize the continuous PDE model in time and space independently.

Time discretization. We define $N_T \in \mathbb{N}^*$, let $\Delta t := T/N_T$ be the constant time-step and $t^k := k\Delta t$, for $k = 0, \dots, N_T - 1$. We consider a partitioning of the time interval $[0,T] = \bigcup_{k=0}^{N_T-1} [t^k, t^{k+1}]$, and we denote by $u^k = u(t^k, x)$ For a general continuous function $u(t,x)$, we approximate the time derivative by

$$\frac{\partial u}{\partial t} \approx \frac{u^{k+1} - u^k}{\Delta t}.$$

In the following, we often use a semi-implicit discretization of the PDE models to simplify our discrete problems. Indeed, some terms are taken at the previous time step to avoid solving nonlinear systems by iterative algorithms. However, a careful discretization must be made to prove that the discrete systems have the same properties as their continuous counterparts. Our goal is to linearize the discrete equations while preserving essential quantities such as preserving mass or the dissipation of the energy associated with the continuous model at the discrete level.

Finite element method. In this manuscript, we mainly focus on the design of efficient numerical methods that use the finite element framework. We aim at propose schemes that can be featured easily in already existing finite element software by making modifications in standard assembling algorithms.

First of all, let us define the spatial discretization of our domain Ω. Let $\mathcal{T}^h$, $h > 0$ be a quasi-uniform mesh of the domain Ω which is defined by disjoint piecewise linear mesh elements, denoted by $T \in \mathcal{T}^h$, such that $\overline{\Omega} \approx \overline{\Omega}_h = \bigcup_{T \in T^h} \overline{T}$. These mesh elements are triangles for $d = 2$ and tetrahedra for $d = 3$. Obviously, it exists other types of element such as quadrangles for $d = 2$, and hexahedra for $d = 3$, however, in this manuscript, we only work with triangular and tetrahedral discretization of Ω. We also denote by N_h, the total number of nodes of $\mathcal{T}^h$, and indicate the set of nodes of $\mathcal{T}^h$ by J_h with $\{x_j\}_{j=1,\dots,N_h}$ the set of their coordinates. If Ω is polyhedral then we have $\Omega = \Omega_h$, however, if for example Ω has a curved boundary, then a small interpolation error is made and $\Omega \neq \Omega_h$.

We let $h := \max_T h_T$ refers to level of refinement of the mesh, where $h_T := \mathrm{diam}(T)$ for $T \in \mathcal{T}^h$, and we define by κ_T the minimal perpendicular length of T and $\kappa_h = \min_{T \in \mathcal{T}^h} \kappa_T$. Furthermore, we assume that the mesh is quasi-uniform, which means it is shape-regular and

there exists a constant $C > 0$ such that

$$h_T \geq Ch, \quad \forall T \in \mathcal{T}^h.$$

We further assume that the mesh is acute, *i.e.* for $d = 2$ the angles of the triangles can not exceed $\frac{\pi}{2}$ and for $d = 3$ the angle between two faces of the same tetrahedron can not exceed $\frac{\pi}{2}$. The acuteness hypothesis is a necessary assumption to have a less restrictive stability criterion for our schemes.

To further improve the stability criteria of the finite element numerical schemes for the PDE models under study, we use a mix of the finite element method with some terms treated as in the finite volume method. To do so, we define the barycentric dual mesh associated to $\mathcal{T}^h$.

First, for an element T with vertices $P_1, \ldots, P_{n_T}$ (n_T being the number of vertices of the element T), we define λ_i (for $i = 1, \ldots, n_T$), the barycentric coordinates of an arbitrary point X inside the element T to be the real numbers satisfying

$$\sum_{i=1}^{n_T} \lambda_i = 1, \quad \text{and} \quad X = \sum_{i=1}^{n_T} \lambda_i P_i.$$

Hence, the barycenter of an element T is the only point for which we have

$$\lambda_i = \lambda_j, \quad \text{for} \quad i = 1, \ldots, n_T, \quad \text{and} \quad \forall j \neq i.$$

From these barycentric coordinates, we can define a subdivision of an element into barycentric subdomains. To denote an element in the domain, we use an index $k \in [1, N_{\text{el}}]$ where N_{el} is the total number of elements in $\mathcal{T}^h$. Therefore, for any element $T_k \in \mathcal{T}^h$, we define the barycentric subdomain associated to the vertex $P_i \in T_k$, by

$$D_i^k := \bigcap_{\substack{j=1 \\ j \neq i}}^{n_T} \{x; x \in T_k \text{ and } \lambda_j(x) \leq \lambda_i(x)\}.$$

Graphically, D_i^k is colored in green in Figure 1.4 for a triangular element. The barycentric dual

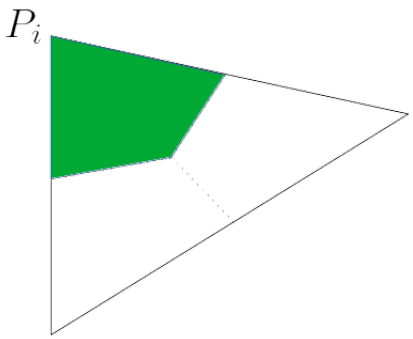


Figure 1.4 – Barycentric subdomain in 2D for the vertex P_i

mesh is defined by these barycentric subdomains. We associate to each node x_i of the mesh $\mathcal{T}^h$ a cell composed of the union of each barycentric subdomains associated to x_i in the elements that share this node. Thus, for each node $x_i \in J_h$, we have the associated cell of the barycentric

dual mesh

$$D_i := \bigcup_k \{D_i^k;\, T_k \in \mathcal{T}^h \text{ such that } x_i \in T^k\}.$$

Figure 1.5 shows an example of triangular mesh for a domain $\Omega \subset \mathbb{R}^2$ and its barycentric dual.

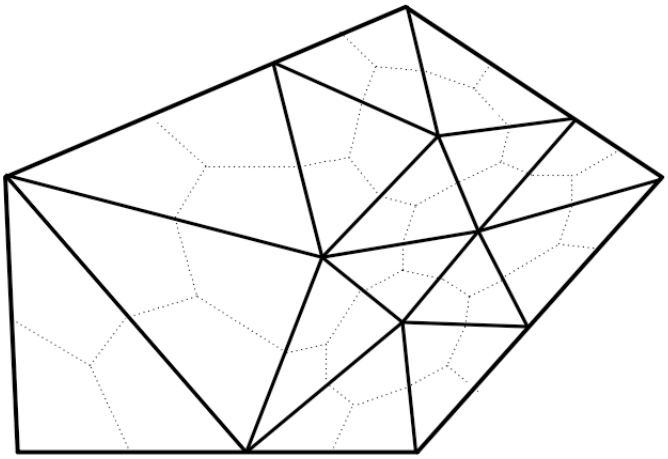


Figure 1.5 – Triangular mesh (solid lines) and its barycentric dual (dotted lines)

We define the P-1 finite element space associated to the mesh $\mathcal{T}^h$ by

$$V_h := \{\chi \in C(\overline{\Omega}) : \chi|_T \in \mathbb{P}^1(T), \quad \forall T \in \mathcal{T}^h\} \subset H^1(\Omega),$$

where $\mathbb{P}^1(T)$ denotes the space of polynomials of order 1 on T. For each node $x_i \in J_h$, we denote by $\chi_i = \chi(x_i)$, the basis function evaluated at the node. Therefore, $\{\chi_j\}_{j=1,\dots,N_h}$ is the standard Lagrangian basis functions associated with the spatial mesh, and we define the approximation of a general function $f \in C^0(\overline{\Omega})$ by $\pi^h f = f_h(x) = \sum_{i=1}^{N_h} f(x_i)\chi_i(x)$.

We combine this basis defined on each element of $\mathcal{T}^h$ with a basis defined on the dual mesh. Let $\hat{\chi}_i \in L^\infty(\Omega)$ be the characteristic function of the barycentric domain D_i associated with each node x_i (for $i = 1, \dots, N_h$). We define the lumped space $\hat{V}_h$ as

$$\hat{V}_h := \{\hat{\chi} :\ \text{piecewise constant over barycentric domains i.e. } \hat{\chi}(x) = \hat{\chi}(x_i),\ \forall x \in D_i\}.$$

This new finite element space is the standard P-0 finite element space on the barycentric dual mesh. We can easily see that the functions $\{\hat{\chi}_j\}_{j=1,\dots,N_h}$ form a basis of $\hat{V}_h$ and we set that they are associative to the functions $\{\chi_j\}_{j=1,\dots,N_h}$ i.e. $\chi(x_i) = \hat{\chi}(x_i)$ for all $x_i \in J_h$.

We also define the lumped scalar product by

$$(v_1, v_2)^h = \int_\Omega \pi^h(v_1(x)v_2(x))\, \mathrm{d}x = (\hat{v}_1, \hat{v}_2)\,, \quad \forall v_1, v_2 \in C^0(\overline{\Omega}),$$

with $\hat{v}_1 = \sum_{x_i \in J_h} v_1(x_i)\hat{\chi}_i$.

The finite element method is applied on the weak formulation of PDEs. Our problem is to find a solution of the PDE model in a certain space V such that its variational form is satisfied for any test function $v \in V$. The Galerkin method consists into approximating the infinite

dimensional space V by V_h of finite dimension. In our case, we use an approximation of the space $H^1(\Omega)$ with V_h, but also an approximation of the space $L^2(\Omega)$ with $\hat{V}_h$.

Using the previous definitions, we can present the standard finite element matrices that are encountered in the manuscript. We denote the standard and lumped mass matrices, associated to the standard L^2 scalar product and to the lumped scalar product, respectively by

$$M_{ij} = \int_\Omega \chi_i \chi_j \, \mathrm{d}x, \quad \text{for } i,j = 1,\dots,N_h,$$

$$M_{l,ij} = \int_\Omega \hat{\chi}_i \hat{\chi}_j \, \mathrm{d}x, \quad \text{for } i,j = 1,\dots,N_h.$$

Therefore, we have that M_l is a diagonal matrix with

$$M_{l,ii} = |D_i| = \frac{1}{3}\left|\left\{\bigcup T \in \mathcal{T}^h \quad \text{such that} \quad x_i \in T\right\}\right|, \quad \text{for } i = 1,\dots,N_h.$$

The stiffness matrix is defined by

$$K_{ij} = \int_\Omega \nabla\chi_i \nabla\chi_j \, \mathrm{d}x, \quad \text{for } i,j = 1,\dots,N_h.$$

Other finite element matrices are defined in the next chapters and are related to specific problems.

Combining the time and space discretizations, we define the approximation of a general function $u(t,x)$ as

$$u(t^k,x) \approx u_h^k(x) = \sum_{i=1}^{N_h} \chi_i(x) u_i^k, \quad \text{or} \quad \hat{u}_h^k(x) = \sum_{i=1}^{N_h} \hat{\chi}_i(x) u_i^k,$$

where the nodal values are defined by $u_i^k = u(t^k, x_i)$.

Upwind method. Let us describe the multi-dimensional upwind approach that we use in the numerical schemes. This method can be easily implemented in standard finite element assembling routines. To explain this upwind method, we use the example of a $\mathbb{P}-1$ finite element method for the first equation of the Keller-Segel model (1.11)

$$\frac{\partial u}{\partial t} = D_u \Delta u - \nabla \cdot (\chi(u) \nabla c),$$

with zero-flux boundary conditions. We assume that the concentration of the chemoattractant is known. In this model, $\chi(u) = \chi_c u (1-u)^\alpha$ is a degenerate chemosensitivity function with $\alpha \geq 0$. A P-1 finite element scheme for this model can be

$$\left(\frac{u_h^{n+1} - u_h^n}{\Delta t}, \chi\right)^h = -D_u \left(\nabla u_h^n, \nabla\chi\right) + \left(\tilde{\chi}(u_h^n)\nabla c_h^n, \nabla\chi\right), \tag{1.13}$$

where $\tilde{\chi}(u_h^n)$ is the upwind mobility that we are going to describe. The key idea of our upwind method is to define an approximation of this mobility function that allows to preserve the non-negativity of the solution. Using the previous structures that we defined above, we approximate the mobility function on each element by a piece-wise constant function $\tilde{\chi}(u_h^n)$. Using the barycentric coordinates, we subdivise each element $T^k \in \mathcal{T}^h$ in $d+1$ subdomains, for $i = 1,2,3$, $j = 2,3$, and $i \neq j$

$$\tilde{D}_{ij}^T = \{x \in T | \lambda_i, \lambda_j \geq \lambda_k, \quad k \neq i,j\}.$$

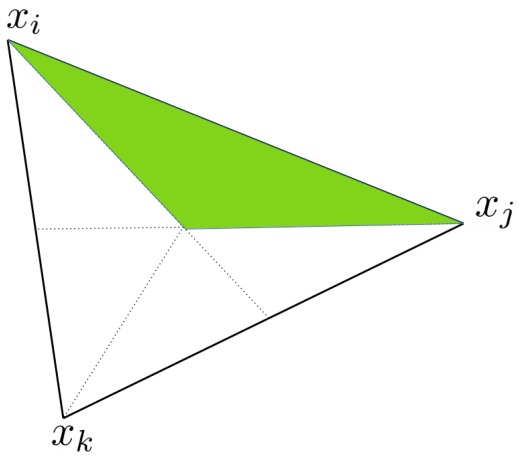


Figure 1.6 – Illustration of the subdomain $\tilde{D}_{ij}^T$ for $d = 2$.

To illustrate what are these subdomains, we represent graphically what is $\tilde{D}_{1,2}^T$ on Figure 1.6 for $d = 2$.

Then, for each $x \in T$, we know that x is in a subdomain $\tilde{D}_{ij}^T$, therefore, for $x \in \tilde{D}_{ij}^T$ we approximate the mobility function by

$$(\tilde{\chi}(u_h^n))(x) = \chi_{ij} = \begin{cases} u_h^n(x_i)(1-u_h^n(x_j))^\alpha, & \text{if} \quad c_h^n(x_i) - c_h^n(x_j) \geq 0, \\ u_h^n(x_j)(1-u_h^n(x_i))^\alpha, & \text{otherwise.} \end{cases}$$

Then, we compute the finite element matrix associated to the right-hand side of (1.13),

$$\begin{aligned} A_{ij} &= \int_\Omega \tilde{\chi}(u_h^n)\nabla\chi_i\nabla\chi_j \, \mathrm{d}x \\ &= \sum_{T\in\mathcal{T}^h} \int_T \tilde{\chi}(u_h^n)\nabla\chi_i\nabla\chi_j \, \mathrm{d}x. \end{aligned}$$

Then, to approximate the last integral, we use a one-point quadrature and choose our quadrature point in the subdomain $\tilde{D}_{ij}^T$ to obtain

$$\int_T \tilde{\chi}(u_h^n)\nabla\chi_i\nabla\chi_j \, \mathrm{d}x \approx \chi_{ij}K_{ij}.$$

As we will see in the following of the manuscript, the combination of the lumping of the mass matrix defined from the barycentric dual mesh and this upwind method for finite element stabilizes the numerical scheme and produces physically relevant solutions for the discrete systems under study.

1.5 Summary of the thesis

In this thesis, we develop, analyze and simulate mathematical models representing pattern formation, and self-organization for cells of living tissues. The manuscript follows a three-part

organization:

— **Part I** is about the Cahn-Hilliard equation for biological application, especially tumor growth. Aiming at designing an efficient structure-preserving numerical method for the Cahn-Hilliard model with degenerate mobility and single-well logarithmic potential, in Chapter 2, we start with a relaxation of the model. This latter reduces to solving of a coupled system of two second-order equations of parabolic and elliptic type. Based on this relaxation, we propose, study and simulate, in Chapter 3, two upwind finite-element schemes for the relaxed degenerate Cahn-Hilliard model.

— **Part II** focuses on deriving new mathematical models for specific applications in Biology. In Chapter 4, we study the effect of the Temozolomide drug on aggregates of glioma cells. We assume that the cells change their mechanical properties in response to the drug, and, as a result, the aggregates shrink. We use a non-linear volume-filling Keller-Segel model to give a mathematical representation of this effect. Then, we simulate our model numerically, and recover a qualitatively good agreement with what is observed in the biological experiments. Then, in Chapter 5, we study the biological scenario of the invasion of a proliferating population of cells in another tissue. We assume that the two populations have different mechanical properties. Since we are convinced that the effects of attraction and repulsion between cells play a crucial role in this scenario, we propose a compressible two-phase Cahn-Hilliard model that is consistent with basic thermodynamics. We also show under which assumptions we can retrieve previous models of the literature that represent this biological scenario.

— **Part III** is about the application of a recent structure-preserving numerical method to non-linear models. In Chapter 6, we use the Scalar Auxiliary Variable method for the simulation of the volume-filling Keller-Segel model. The complete numerical scheme combines both the upwind method inside the finite element framework that we propose in Chapter 3, and the SAV method. We explore the analytical properties of this numerical scheme and present some numerical simulations that seem to indicate a potential advantage of the SAV method compared to other numerical methods. To investigate the changes induced by the SAV method, in Chapter 7, we apply the method to a simpler model that arise in Quantum Physics. After reviewing the critical analytical properties of the scheme, we present a comparison between the SAV methods and numerical methods of reference for this problem. This allows us to discuss the advantages of the SAV method for specific applications.

In the following of this section, we summarize the main contributions of our work.

1.5.1 Towards an efficient numerical scheme for the degenerate Cahn-Hilliard model for Biology

The degenerate Cahn-Hilliard model (1.4) with a single-well logarithmic potential induces numerous difficulties both at the analytical and at the numerical level. Indeed, for our order parameter $n \in [0, 1)$, degeneracy of the mobility coefficient in $n = 0$ and $n = 1$ as well as the singularity of the potential $\psi(1) = +\infty$ makes the theoretical analysis substantially harder compared to the constant mobility case with a smooth polynomial potential. Furthermore, the fact that the degeneracy and singularity sets do not coincide, i.e. the single-well potential is degenerate in $n = 0$ but not singular, produces instabilities at the discrete level. Indeed, the solution of a standard discretization of the DCH model with single-well potential can become negative, which is not physically consistent. Our work aims at solving this issue while designing an efficient numerical scheme.

Relaxation of the degenerate Cahn-Hilliard model

Based on analytical results on the degenerate Cahn-Hilliard with double-well singular potential [76], Agosti *et. al.* [8] analyzed the model properties when the potential is single-welled and logarithmic. The authors show the existence of global weak solutions from compactness properties and energy apriori estimates. Based on this work and aiming to design an efficient numerical scheme for this model, we propose modifying the equation. A common approach to reduce the difficulty induced by the equation order is to split the equation into a system of two second-order equations. Based on this idea and using the fact that the single-well potential can be decomposed as the sum of a convex and a non-convex part, we propose the model

$$\begin{cases} \partial_t n_\sigma = \nabla \cdot \left(b(n_\sigma) \nabla \left(\varphi_\sigma + \psi_+'(n_\sigma) \right) \right) & \text{in } \Omega \times (0, +\infty), \\ -\sigma \Delta \varphi_\sigma + \varphi_\sigma = -\gamma \Delta n_\sigma + \psi_-' \left(n_\sigma - \dfrac{\sigma}{\gamma} \varphi_\sigma \right) & \text{in } \Omega \times (0, +\infty), \end{cases} \tag{1.14}$$

that we complete with zero-flux boundary conditions

$$\frac{\partial(\gamma n_\sigma - \sigma \varphi_\sigma)}{\partial \nu} = b(n_\sigma) \frac{\partial \left(\varphi_\sigma + \psi_+'(n_\sigma) \right)}{\partial \nu} = 0 \qquad \text{on } \partial\Omega \times (0, +\infty). \tag{1.15}$$

In this form, the backward diffusion (aggregation part of the potential) is contained inside the new variable φ. We further added a diffusion term in the equation for this new variable to regularize the model. As a result, the original fourth-order equation is transformed into a system of two second-order equations that features an extra regularization in space of the variable containing the unstable term of backward diffusion. The decomposition of the potential in a convex and a non-convex part will also be useful for the design of our numerical scheme. This idea originates from the work of Eyre [82].

In Chapter 2 and [165], we are interested in proving the existence of global weak solutions $\{n_\sigma, \varphi_\sigma\}$ for (1.14), and also check if we can prove that we recover the original Cahn-Hilliard model in the limit $\sigma \to 0$. We investigate also the behavior in the long-time of the solutions and compare it to the one of the original equation. All the proofs in this work rely on apriori estimates and compactness results. Mainly, an energy associated to the system (1.14) with boundary conditions (1.15) can be found

$$\mathcal{E}_\sigma[n_\sigma] = \int_\Omega \left[\psi_+(n_\sigma) + \frac{\gamma}{2} |\nabla (n_\sigma - \frac{\sigma}{\gamma} \varphi_\sigma)|^2 + \frac{\sigma}{2\gamma} |\varphi_\sigma|^2 + \psi_-(n_\sigma - \frac{\sigma}{\gamma} \varphi_\sigma) \right],$$

and we also know that it dissipates i.e.

$$\frac{d}{dt} \mathcal{E}_\sigma[n_\sigma(t)] = - \int_\Omega b(n_\sigma) \left| \nabla (\varphi_\sigma + \psi_+'(n_\sigma)) \right|^2 \leqslant 0.$$

However, from the energy, we cannot find an apriori estimate on $\nabla \varphi_\sigma$. To tackle this issue, we calculate the entropy of the system. Indeed, definition a convex function $\phi(\cdot)$ such that $\phi''(n_\sigma) = \frac{1}{b(n_\sigma)}$, we find the entropy estimate

$$\begin{aligned} \frac{d\Phi[n_\sigma(t)]}{dt} = - \int_\Omega \gamma \left| \Delta \left(n_\sigma - \frac{\sigma}{\gamma} \varphi_\sigma \right) \right|^2 + \frac{\sigma}{\gamma} |\nabla \varphi_\sigma|^2 + \psi_-''(n_\sigma - \frac{\sigma}{\gamma} \varphi_\sigma) \left| \nabla (n_\sigma - \frac{\sigma}{\gamma} \varphi_\sigma) \right|^2 \\ + \psi_+''(n_\sigma) |\nabla n_\sigma|^2. \end{aligned}$$

To show the existence of global weak solutions for the relaxed problem (1.15), we use a

regularization method. Indeed, using a positive parameter ε, we define a positive mobility and a non-singular potential. On this regularized model, we are able to compute the same energy and entropy functionals. Then, from these two apriori estimates, and due to the fact that they are bounded uniformly in ε, we are able to show the existence of global weak solutions for model (1.14)–(1.15) from compactness results (Proposition 5). We also show that these weak solutions satisfy $0 \leq n_\sigma < 1$. The proof of this result is based on a contradictory argument obtained from the boundedness of the entropy. Indeed, since in the case of the degenerate mobility $b(n_\sigma) = n_\sigma(1-n_\sigma)^2$, the entropy behaves as

$$\begin{aligned} \phi(n_\sigma) &= n_\sigma \log(n_\sigma), \quad \text{for} \quad n_\sigma \approx 0^+, \\ \phi(n_\sigma) &= -\log(1-n_\sigma), \quad \text{for} \quad n_\sigma \approx 1^-, \end{aligned}$$

and, assuming proper initial conditions in $H^1(\Omega)$, we know that the entropy functional $\Phi[n_\sigma(t,x)] = \int_\Omega \phi(n_\sigma(t))\, \mathrm{d}x$ is bounded, and we conclude that the solutions n_σ remain in $[0,1)$. Using the fact that entropy and energy estimates are uniformly bounded in σ, we are able to present the main theorem of Chapter 2

Theorem 6 (Limit $\sigma = 0$) *Let $(n_{\sigma,\epsilon}, \varphi_{\sigma,\epsilon})$ be a sequence of weak solutions of the regularized-relaxed degenerate CH model. Then, assuming initial conditions $\{u^0, \varphi^0\} \in H^1(\Omega) \times H^1(\Omega)$ with $0 \leq n^0 < 1$, we can extract a subsequence of $(n_{\sigma,\epsilon}, \varphi_{\sigma,\epsilon})$ as $\epsilon,\ \sigma \to 0$, such that*

$$\varphi_{\sigma,\epsilon} \rightharpoonup -\gamma\Delta n + \psi'_-(n) \qquad \textit{weakly in } \ L^2(\Omega_T),$$

$$n_{\sigma,\gamma} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon} \to n \qquad \textit{strongly in } \ L^2(0,T;H^1(\Omega)),$$

$$n_{\sigma,\epsilon},\ \nabla n_{\sigma,\epsilon} \to n,\ \nabla n \qquad \textit{strongly in } \ L^2(\Omega_T), \ \textit{and} \ \ 0 \leqslant n < 1,$$

and $n < 1$ a.e. if b vanishes fast enough at 1 so that $\phi(1) = \infty$.

$$\partial_t n_{\sigma,\epsilon} \rightharpoonup \partial_t n \ \textit{weakly in } L^2\big(0,T;(H^1(\Omega))'\big).$$

The limit n satisfies the DCH system (1.4) in the weak sense.

Therefore, from this result, we can confirm that the relaxed-degenerate Cahn-Hilliard model (RDCH in short) is good approximation of the original DCH model. We propose a proof of that in the complicated context of the single-well logarithmic potential. Obviously, this result also extends to the case of double-well smooth and logarithmic potentials.

Then, in the end of Chapter 2, based on the control provided by energy and entropy estimates, we explore the behavior of the solutions of the RDCH in the long-time. Defining $n_k(t,x) = n(t+k,x)$, and $\varphi_k(t,x) = \varphi(t+k,x)$, we explore the large time limit $k \to \infty$. Numerically, we observe that the long-time behavior of the solutions of the RDCH model meets the analytical description of the steady-states given by Songmu [183]: n varies smoothly in space from plateaus of maximum value $n = 1$ to zones where $n = 0$ (see Figure 1.7). Analytically, we retrieve the convergence of the solutions of the RDCH model to steady-states

Proposition 7 (Long term convergence along subsequences) *Let (n, φ) be a weak solution of the RDCH model with initial condition n^0 with $0 \leq n^0 < 1$, and finite energy and entropy. Then, we can extract a subsequence, still denoted by index k, of (n_k, φ_k) such that*

$$\lim_{k\to\infty} n_k(x,t) = n_\infty(x), \quad \lim_{k\to\infty} \varphi_k(x,t) = \varphi_\infty(x) \qquad \textit{strongly in } L^2\big((-T,T)\times\Omega\big), \quad \forall T > 0,$$

where $(n_\infty, \varphi_\infty)$ *are solutions of* (2.83) *satisfying*

$$b(n_\infty)\nabla\left(\varphi_\infty + \psi'_+(n_\infty)\right) = 0.$$

Lastly, we conclude Chapter 2 by arguing that the RDCH model can be easily implemented in standard finite element software, the only difficulty to be tackled being the loss of the positivity of n at the discrete level.

Upwind finite-element scheme

Based on the work of Barrett *et. al.* [28], Agosti *et. al.* [8] proposed to use a finite element scheme to simulate the DCH model that solves a variational inequality in order to enforce positivity of the relative cell density n and the dissipation of the energy at the discrete level. However, this non-linear numerical scheme is rather complicated to use since it requires an iterative algorithm and a slight twist of the total mass to preserve the other quantities. Furthermore, the scheme requires a lot of computational power. Indeed, even though the scheme is non-linear, it is not entirely implicit and suffers restrictions for choosing the time step to remain stable and accurate. To tackle the non-preservation of the initial mass, Agosti [7] proposed a discontinuous Galerkin discretization of the DCH model. However, this has not solved the computational cost issue since the discontinuous Galerkin method is well-known to be a computationally expensive method. However, the numerical simulations conducted from [8] were applied in a concrete study with patient data [10] showing a good agreement in evolution and volume of the tumor even considering the effect of treatment in the model.

Based on the relaxation of the model made in Chapter 2, and to tackle some issues experienced with previous numerical schemes for the DCH model, we conduct in Chapter 3 the design of a P-1 finite element scheme for the relaxed-degenerate Cahn-Hilliard model. We use the finite element framework presented in the previous section. Two different classes of numerical schemes are presented in this chapter, the main difference between the two being the approximation of the continuous mobility function b. In the first class of schemes, we follow the idea of Grün and Rumpf [108] and use a piecewise constant matrix defined for each element of the mesh to approximate the mobility. The advantage of this method is that it allows to obtain an entropy estimate at the discrete level. However, this method requires the mesh to be composed of right-angled elements for $d = 2, 3$. For this method, the solutions of the discrete scheme converge to the solutions of the continuous RDCH model as $\Delta t, h \to 0$. To perform the convergence analysis, we used the discrete version of the entropy estimate since the mobility is defined to satisfy the Definition 8. Then, we can use the projection of the derivative of the entropy on V^h to compute the entropy functional.

Definition 8 (mobility-entropy pair) *An admissible entropy-mobility pair* $\{M, \phi\}$ *with respect to the triangulation* $\mathcal{T}^h$ *satisfies the following axioms*

i) $M : V^h \to \otimes_{k=1}^{|\mathcal{T}^h|}\mathbb{R}^{d\times d}$ *is continuous;*

ii) $M(s)\big|_K = b(s)I_d$ *if* s *is constant on the element* $K \in \mathcal{T}^h$*;*

iii) $M^T(s)\nabla\pi^h(\phi'(s)) = \nabla s$*;*

iv) on each element $K \in \mathcal{T}^h$*, the matrix* $M(s)\big|_K$ *is symmetric and positive semidefinite.*

Then, from the energy and entropy estimates obtained on the regularized discrete problem and using compactness results, we are able to show the existence of a solution $\{n_h^{k+1}, \varphi_h^{k+1}\} \in V_h \times V_h$ with $0 \leq n_h^{k+1} < 1$ for the non-regularized problem, and to perform a convergence analysis.

To relax the previous constraint on the mesh (i.e. the mesh is composed of right-angled elements), we propose an adaptation of the upwind method to compute the mobility at the discrete level. This method allows to preserve the physical bounds of the cell density and dissipate the energy for a non-linear semi-implicit time discretization.

The upwind method plays a role only in the calculation of the non-constant finite element matrix U associated with the right-hand side of the first equation of the model (1.14). From our multi-dimensional upwind method, we know that each entry of this matrix is given for $i, j = 1, \dots, N_h$, by

$$\begin{aligned} U_{ij} &= \int_\Omega \tilde{b}(n_h^{k+1}) \nabla \chi_i \nabla \chi_j \, \mathrm{d}x, \\ &\approx B_{ij}^{k+1} \int_\Omega \nabla \chi_i \nabla \chi_j \, \mathrm{d}x. \end{aligned}$$

This upwind mobility $\tilde{b}(n_h^{k+1})$ is defined for every pair of nodes $\{x_i, x_j\}$ as a constant in the domain $\tilde{D}_{ij}^T$ defined in the section 1.4. Then, as in the finite volume method, defining

$$\xi_i^{k+1} := \left(\varphi_h^{k+1} + \psi_+'(n_h^{k+1})\right)(x_i),$$

which is constant on each cell of the barycentric dual mesh D_i, we look at the sign of the difference $\xi_j^{k+1} - \xi_i^{k+1}$. This sign give us an information on the direction of the transport. Therefore, inspired by the calculation of the upwind chemosensitivity in [45], we propose to take

$$B_{ij}^{k+1} := \begin{cases} n_i^{k+1}(1 - n_j^{k+1})^2, & \text{if } \xi_i^{k+1} - \xi_j^{k+1} > 0, \\ n_j^{k+1}(1 - n_i^{k+1})^2, & \text{otherwise,} \end{cases} \qquad i, j = 1, \dots, N_h.$$

In Chapter 3, we propose two schemes based on two different implicit-explicit time discretizations: a nonlinear scheme

$$\begin{cases} \left(\frac{n_h^{k+1} - n_h^k}{\Delta t}, \chi\right)^h + \left(\tilde{b}(n_h^{k+1}) \nabla \left(\varphi_h^{k+1} + \pi^h(\psi_+'(n_h^{k+1}))\right), \nabla \chi\right) = 0, \\ \sigma \left(\nabla \varphi_h^{k+1}, \nabla \chi\right) + \left(\varphi_h^{k+1}, \chi\right)^h = \gamma \left(\nabla n_h^{k+1}, \nabla \chi\right) + \left(\psi_-'(n_h^k - \frac{\sigma}{\gamma} \varphi_h^k), \chi\right)^h, \end{cases}$$

and a linear one

$$\begin{cases} \left(\frac{n_h^{k+1} - n_h^k}{\Delta t}, \chi\right)^h + \left(b(n_h^k) \psi_+''(n_h^k) \nabla n_h^{k+1}, \nabla \chi\right) = -\left(\tilde{b}(n_h^k) \nabla \varphi_h^{k+1}, \nabla \chi\right), \\ \sigma \left(\nabla \varphi_h^{k+1}, \nabla \chi\right) + \left(\varphi_h^{k+1}, \chi\right)^h = \gamma \left(\nabla n_h^k, \nabla \chi\right) + \left(\psi_-'(n_h^k - \frac{\sigma}{\gamma} \varphi_h^k), \chi\right)^h. \end{cases}$$

For the first, we are able to prove that this scheme preserves the dissipation of the energy at the discrete level, and we can derive an energy estimate. However, compared to the continuous case, we can not define the entropy in a standard way at the discrete level and control the last terms that we needed to make a convergence analysis.

For the practical and efficient linear scheme that we use for the numerical simulations, we can show that it admits a unique solution and preserves the positivity of n_h^{k+1}. Then, we present 1D and 2D numerical simulations and compare them to the numerical experiments made by Agosti [8]. We observe that even if we are simulating the RDCH model instead of the original DCH model, the solutions given by the simulations are comparable. We cannot prove analytically that the linear scheme dissipates the energy. However, we can see that this property is satisfied during the simulations. Then, to explain the advantage of our relaxation that we have made at

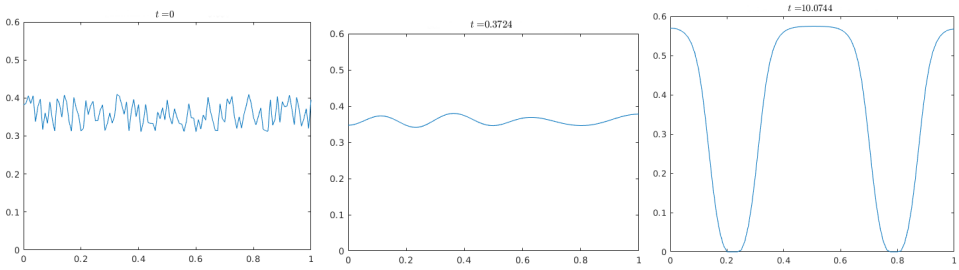


Figure 1.7 – Numerical simulation for $d = 1$ of the RDCH model. From the initial condition (left), through the first stage (middle), to the stable steady-state (right).

the continuous level, we test the stability of our scheme for various values of σ by calculating the spectral radius of the amplification matrix, and we observe that the relaxation allows to take larger time steps.

1.5.2 Modelling of specific scenarios in Biology

Undertaking numerous biological experiments to unveil the mechanisms that cells of tissues use in specific processes is costly in terms of time and money. Mathematical models provide a framework in which different hypotheses can be tested and compared to the experiments. If a model reproduces the behavior of the cells well, it gives an idea in which direction the biologists may orient their research. The Cahn-Hilliard equation and the Keller-Segel model are general models in which more physical effects can be added to represent particular scenarios in Biology. In this manuscript, we present two mathematical models that can help biologists since the numerical simulations are qualitatively in good agreement with the experiments.

Our first study comes from a collaboration with biologists from the Centre de Recherche en Cancérologie et Immunologie Nantes-Angers to understand the possible change in mechanical properties of glioblastoma cells in response to a chemotherapeutic treatment. The second one aims at understanding the role of multiple mechanical effects such as viscosity or friction in the unstable invasion of tumor cells in healthy tissue.

Change in mechanical properties of tumor cells due to treatment

From the biological observation during in-vitro experiments that spheroids of glioblastoma cells shrink in response to chemotherapeutic treatment, biologists wonder what cells are undergoing in this process. In Chapter 4, we assume that spheroids of glioblastoma cells are formed due to chemotaxis. This hypothesis was studied by mathematicians in [45], and is supported by biologists since the tumor cells release chemical components, such as cytokines, in their microenvironment known to drive the movement of cells. To understand the shrinking of these in-vitro aggregates in response to a particular concentration of the chemotherapeutic drug Temozolomide (TMZ), we propose to study the hypothesis that cells change their mechanical properties. When the concentration of TMZ increases, the tumor cells switch from hard spheres to semi-elastic bodies, and, therefore, in the presence of TMZ, tumor cells can squeeze. As a consequence, the spheroids decrease in volume, i.e., the aggregates shrink.

To represent this effect in a mathematical model, we use the framework provided by the nonlinear volume-filling Keller-Segel model, and modify both the chemosensitivity and the diffusion terms associated to the cells. We propose a two-part modelling: in the first part, the TMZ is not introduced yet, and the model is the standard volume-filling Keller-Segel model with a logistic

proliferation term for the cells, whereas, for the second part, we add a supplementary diffusion equation for the TMZ that is coupled in a nonlinear manner to the equation for the cells.

Let us consider a domain with a smooth boundary (see Definition 1). For $u(t,x)$ the density of glioblastoma cells, $c(t,x)$ the concentration of chemoattractant and $M(t,x)$ the concentraton of TMZ, the model for the first part of our study is

$$\begin{cases} \partial_t u &= \nabla \cdot (d_1 \nabla u - \chi_u \phi_1(u) \nabla c) + f(u) \ , \\ \partial_t c &= d_2 \Delta c + \alpha u - \beta c \ , \end{cases} \tag{1.16}$$

and for the second part (i.e. the effect of TMZ on already formed spheroids) is

$$\begin{cases} \partial_t u &= \nabla \cdot (d_3 D_2(u,M) \nabla u - \chi_u \phi_2(u,M) \nabla c) + f(u) \ , \\ \partial_t c &= d_2 \Delta c + \alpha u - \beta c \ , \\ \partial_t M &= d_4 \Delta M - \delta u \ , \end{cases} \tag{1.17}$$

where $d_1, d_3, d_2, d_4, \chi_u, \alpha, \beta, \delta$ are constant parameters, the chemosensitivity for the first part is $\phi_1(u) = u(1 - u/u_{\max})$ and $f(u)$ is a logistic growth function

$$f(u) = r_0 u \left(1 - \frac{u}{u_{\max}}\right),$$

with $u_{\max}$ denoting the carrying capacity i.e. the critical value of cell density above which cells do not proliferate anymore. This model is supplemented by zero-flux boundary conditions i.e. the walls of the domain are not permeable. The functions related to the non-constant part of the diffusion coefficient, and to the strength of the chemotactic effect are given respectively by

$$D_2(u,M) = 1 + (\gamma_M - 1)\left(\frac{u}{\overline{u}}\right)^{\gamma_M} \quad \text{and} \quad \phi_2(u,M) = u\left(1 - \left(\frac{u}{\overline{u}}\right)^{\gamma_M}\right),$$

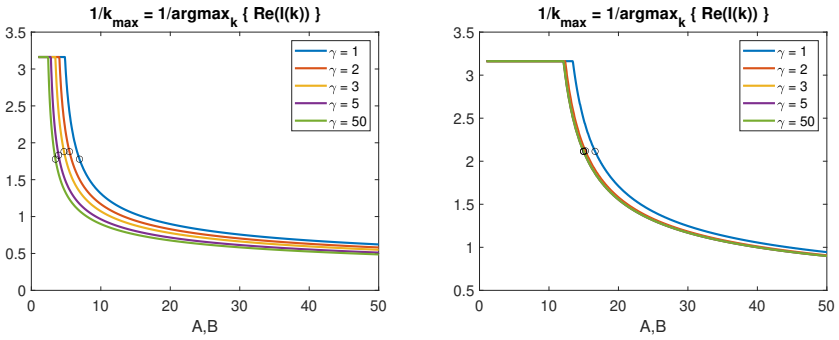
with $\overline{u} \geq u_{\max}$ being the packing capacity above which cells repel each other to avoid overcrowding. We denote by $\gamma_M \geq 1$ the squeezing parameter, which is a function of the TMZ concentration M. This term is related to the elasticity of the cells. The larger γ_M is, the more cells are elastic.

To understand the patterns we can expect from different values for the parameters, we conduct a linear stability analysis. This analysis gives what value should take the ratio between the chemotactic strength and the diffusion of the cells to see the emergence of aggregates as a function of the other parameters that drive the cells' proliferation and packing capacity. On Figure 1.8, we present for different values of r_0 and γ_M, what are the values of

$$A = \frac{\chi_u}{d_1}, \quad \text{and} \quad B = \frac{\chi_u}{d_3},$$

such that, we see the emergence of patterns. This formation of structures is observed when $\frac{1}{k_{\max}} \leq k_c$, where $k_{\max}$ is the maximal wavefunction and k_c the critical value such that the perturbed linearized system is unstable.

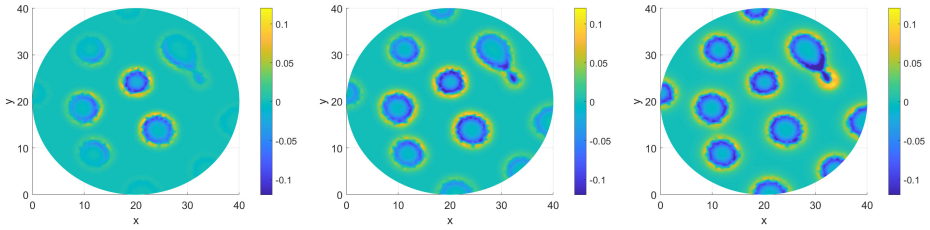
After this linear stability analysis, we present numerical simulations based on the numerical scheme presented in Chapter 3. We explore the effect of the length of the domain through different scenarios such as introducing the drug at different times or for different initial conditions for M. Then, we show the shrinking of the aggregates due to the TMZ in 2D numerical simulations. Indeed, considering a Gaussian located at the center of a circular domain Ω as the initial condition for M, and the initial condition for the cell density and chemoattractant being the result of the



(a) (b)

Figure 1.8 – Wavenumbers for different values of γ_M when (a) $r_0 = 0.1$, $u_{\max} = u_0 = 0.5$; and for (b) $r_0 = 0.05$, $u_{\max} = u_0 = 0.1$. Circles indicate the values of k_c.

simulation of the first model (without TMZ), we observe the evolution depicted in Figure 1.9. This figure shows the difference between the initial condition for u and the new cell density at different times. We observe that the aggregates are shrinking, and this numerical result is in qualitatively good agreement with the biological experiments. Therefore, based on these numerical results, we propose that the reason for the shrinking of the aggregates of glioblastoma cells in response to TMZ is the changes of mechanical properties of the cells.



(a) (b) (c)

Figure 1.9 – Difference between the solutions when the initial concentration of the treatment is a Gaussian function centered in the domain. (a) $T = 210$, (b) $T = 230$ (c) $T = 300$ for $T_1 = 200$, $r_0 = 0.05$ and $u_0 = 0.1$.

Instabilities at the border of invasive tumors

The invasion of a tumor in healthy tissue is often associated with the emergence of irregularities at the tumor surface. To understand the relevant physical effects playing a role in these structural instabilities, we propose a general model based on the theory of mixture. We propose to represent two populations of cells with different properties using a mixture composed of two different components. We also specify that one population is proliferating while the other is not. Let population 1 be the proliferating population, and let us denote ρ_i the relative density of the i-th component, $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$ the velocity of the mixture, c the mass fraction of component 1.

We assume the following:

— the mixture is compressible, and the proliferation of the cells $G(\cdot)$ depends on the pressure inside the mixture, that we denote in our model by p;

— the mixture is viscous and it exists an interfacial force between the two populations where the width of the diffuse interface is proportional to $\sqrt{\gamma}$;

— to take into account the role of the extracellular matrix, which is assumed to be a lattice of stiff fibers, we consider the velocity of the mixture to depend on a friction force $\kappa(\cdot)$, which can be different in function of the cell type;

— cells exert attraction and repulsion forces depending on their type and relative density, modeled by the potential ψ_0.

Therefore, from general conservation laws for the two components of the mixture, and using basic mechanics, we derive a generalized Navier-Stokes-Cahn-Hilliard model (generalized NSCH in short)

$$\frac{\partial \rho}{\partial t} + \mathrm{div}\,(\rho \mathbf{v}) = \rho c G(p), \quad \text{(continuity equation for total density)},$$

$$\rho \frac{\mathrm{D}c}{\mathrm{D}t} = \mathrm{div}\,(b(c)\nabla \mu) + \rho c(1-c)G(p), \quad \text{(continuity equation for mass fraction of component 1)},$$

$$\rho \mu = -\gamma \mathrm{div}\,(\rho \nabla c) + \rho \frac{\partial \psi_0}{\partial c}, \quad \text{(definition of chemical potential)},$$

$$\begin{aligned} \rho \frac{\mathrm{D}\mathbf{v}}{\mathrm{D}t} =& -\left[\nabla p + \gamma \mathrm{div}\,(\rho \nabla c \otimes \nabla c)\right] + \mathrm{div}\left(\nu(c)\left(\nabla \mathbf{v} + \nabla \mathbf{v}^T\right)\right) \\ & - \frac{2}{3}\nabla\left(\nu(c)\left(\mathrm{div}\,(\mathbf{v})\right)\right) - \kappa(c)\mathbf{v} - \rho c \mathbf{v} G(p), \quad \text{(equation for the velocity of the fluid)} \end{aligned}$$

supplemented by homogeneous Neumann boundary conditions

$$\frac{\partial \mu}{\partial \mathbf{n}} = \frac{\partial \mathbf{v}}{\partial \mathbf{n}} = \frac{\partial \rho c}{\partial \mathbf{n}} = \frac{\partial \rho}{\partial \mathbf{n}} = 0,$$

where $\mathbf{n}$ is the outward normal vector to the boundary $\partial\Omega$. From a set of constraints based on thermodynamics laws, we derive a physically relevant model. Altogether, this model is a generalization of previous researches about the compressible Cahn-Hilliard model. However, this kind of models has never been applied to biological applications.

To understand the link with previous mathematical representations of invasion from the literature, we change our model using simplifying assumptions. Mainly, we assume that friction on the extra-cellular matrix is the preponderant effect. By formal asymptotical limits, we recover models of the type of that considered in [139], which describes the dynamics of two populations of cells with different mobilities and different proliferation rates

$$\begin{cases} \frac{\partial \rho_1}{\partial t} - \mu_1 \mathrm{div}\,(\rho_1 \nabla p) = \rho_1 G(p), \\ \frac{\partial \rho_2}{\partial t} - \mu_2 \mathrm{div}\,(\rho_2 \nabla p) = 0. \end{cases}$$

Here, ρ_1 and ρ_2 are the densities of the two populations, and μ_1 and μ_2 are the different mobility coefficients of cells in the two populations. Since $p = (\rho_1 + \rho_2)^\alpha$ ($\alpha \geq 1$), the two equations are nonlinear and coupled. Numerical simulations of the this model show the formation of finger-like instabilities if the proliferating population is moving faster (see Figure 1.10 where we reproduce the results of Lorenzi *et. al.* [139]).

Our study indicates that the model from [139] is physically relevant even though it was presented as a phenomenological representation of two cell populations. Indeed, in a certain regime, our NSCH model, which is fully consistent with thermodynamics and mechanics laws,

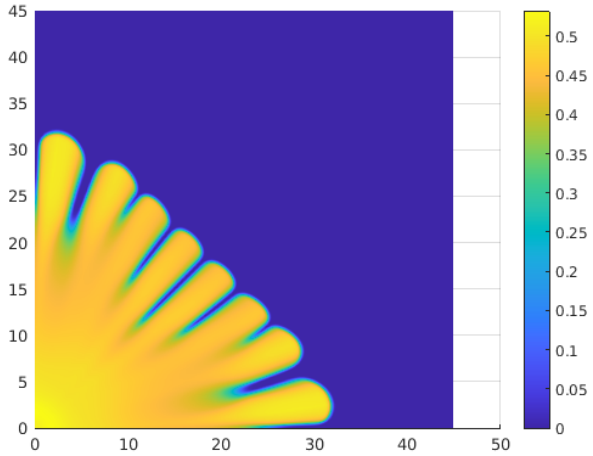


Figure 1.10 – Numerical simulation of the model from [139] with $\mu_1 > \mu_2$.

converges to the model (1.5.2). This asymptotic limit is obtained assuming that the friction of cells on the ECM is the predominant effect. We plan to explain the role of the other effects such as attraction-repulsion and viscosity of the fluids (with a possible contrast for the two components) using numerical simulations. Our ultimate goal is to give a list of the possible physical effects that play a role in the formation of instabilities at the surface of a growing tumor.

1.5.3 Structure-preserving numerical method for nonlinear models

Preserving the properties of the continuous form of the PDE models for their numerical schemes often requires an implicit discretization since it is closer in some sense to the continuous equation. However, for nonlinear PDE models, the preserving of the properties at the discrete level requires to solve nonlinear ODE systems and, hence, to use costly iterative solver. Even though implicit methods allow for larger time steps, one must often take a fine temporal grid for reasons of accuracy. Therefore, in Part III we aim to find a structure-preserving method for living tissue PDE models that allows for an implicit-explicit discretization such that the resulting scheme is linear. To solve this problem, Shen *et. al.* [181, 179, 180] proposed a method initially designed for nonlinear equations with a gradient-flow structure: the Scalar Auxiliary Variable method. This method is designed to treat the nonlinear terms explicitly in the discrete model. Even though the resulting scheme is linear, we are able to prove analytically that it dissipates a modified energy. However, this method has been applied only on models involving constant coefficients and models that do not induce instabilities at the discrete level.

Numerical method for the nonlinear volume-filling Keller-Segel model

In Chapter 6, we apply the SAV method to design a structure-preserving method for the volume-filling KS model (1.11). This model is nonlinear and has a gradient flow structure. Indeed, for $u(t,x)$ the cell density, and $c(t,x)$ the chemoattractant produced by the cells, the

energy is given by

$$\mathcal{E}[u,c](t) = \int_\Omega \frac{D_u}{\chi_c}\left[u\log u - (u-1)\log(1-u)\right] - uc + \frac{1}{2}\left(|\nabla c|^2 + \alpha c^2\right) + C \ \mathrm{d}x,$$

with $C \geq 0$ a constant such that the energy is positive. The gradient-flow form of the model is then given by

$$\begin{cases} \partial_t u &= \nabla\cdot\left(\chi_c\varphi(u)\nabla\frac{\delta\mathcal{E}}{\delta u}\right), \\ \tau\partial_t c &= -\frac{\delta\mathcal{E}}{\delta c}. \end{cases}$$

The SAV method introduces a new unknown $r = \sqrt{\mathcal{E}_1[u]}$ that "hides" the nonlinear terms such that

$$\begin{cases} \partial_t u &= \nabla\cdot\left(\chi_c\varphi(u)\nabla\mu_1\right), \\ \mu_1 &= B\frac{r}{\sqrt{\mathcal{E}_1[u]}}g(u) - c, \\ \tau\partial_t c &= -\mu_2, \\ \mu_2 &= -\Delta c + \alpha c - u, \\ \frac{\mathrm{d}r}{\mathrm{d}t} &= \frac{1}{2\sqrt{\mathcal{E}_1[u]}}\int_\Omega g(u)\frac{\partial u}{\partial t}\ \mathrm{d}x, \end{cases}$$

is equivalent to the previous system.

From this SAV-KS model, we apply a P-1 finite element method and stabilize the scheme using the multi-dimensional upwind method designed for the relaxed-degenerate Cahn-Hilliard model. Altogether, our numerical scheme is an implicit-explicit linear finite element discretization of the volume-filling KS model. We are able to show the existence of a unique solution to the discrete system. Using results from mass-lumping stabilization of finite element discretization for parabolic equation [94], we also show that our scheme preserves the positivity of the cell density u, and its upper bound. Since we used the SAV method, we prove analytically that the scheme dissipates a modified version of the energy.

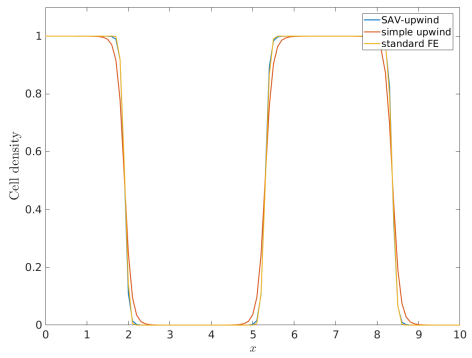
An interesting observation is made while presenting our numerical results: the combination of the SAV method with the multi-dimensional upwind method enhances the spatial order of accuracy. Indeed, it is well-known that using upwinding, we expect to recover a method that is less than first-order accurate in space. Computing the order of accuracy numerically, we recover a slope between first and second order. Comparing the results obtained with a standard finite element discretization of the scheme, a classical upwind method, and our SAV-upwind numerical scheme, we observe that our scheme can remain accurate even in sharp zones (see Figure 1.11).

Long-time behavior of the Scalar Auxiliary Variable method for dispersive equations

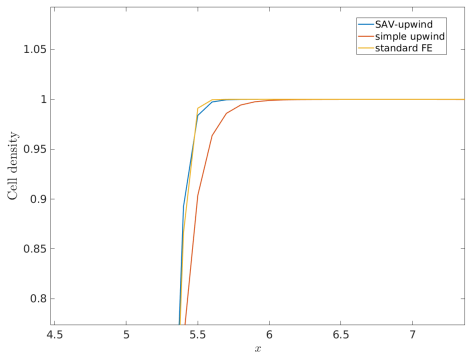
Understanding the properties of numerical schemes designed from the Scalar Auxiliary Variable remains unclear. Several applications of this method have been proposed for nonlinear gradient-flow models and dispersive equations. For these latter, the conservation properties are now well understood. A SAV scheme for a nonlinear dispersive equation allows for an unconditional conservation of a modified Hamiltonian energy. However, for the Nonlinear Schrödinger (NLS in short) equation, it remains unclear how the error on the real Hamiltonian evolves, especially for longtime simulations. For a smooth domain Ω, the NLS equation reads

$$i\partial_t u(t,x) = -\Delta u(t,x) + V(x)u(t,x) + f\left(|u(t,x)|^2\right)u(t,x), \quad t\in(0,T], \quad x\in\Omega$$

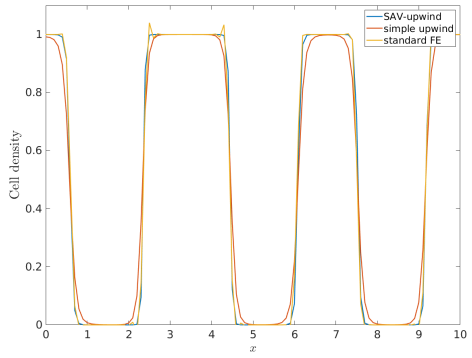
In Chapter 7, we propose a SAV numerical scheme for the NLS equation using its Hamiltonian formulation. We use a Crank-Nicholson temporal discretization, and a Fourier pseudo-spectral



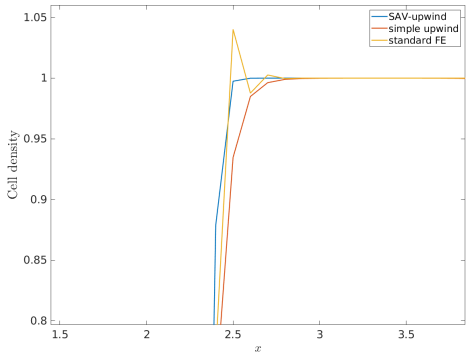
(a) Stable case



(b) Top of an aggregate for stable case



(c) Unstable case



(d) Top of an aggregate for unstable case

Figure 1.11 – End state of the simulations of the volume-filling KS model for a standard finite element discretization, a classical upwind method in finite element, the SAV-upwind method.

method for the space discretization. Our scheme reads

$$\begin{cases} \frac{p^{k+1}-p^k}{\Delta t} & = -\Delta q^{k+1/2} + r^{k+1/2}\tilde{g}_1^{k+1/2}, \\ \frac{q^{k+1}-q^k}{\Delta t} & = \Delta p^{k+1/2} - r^{k+1/2}\tilde{g}_2^{k+1/2}, \\ r^{k+1} - r^k & = \frac{1}{2}\left[\left(\tilde{g}_1^{k+1/2}, q^{k+1}-q^k\right) + \left(\tilde{g}_2^{k+1/2}, p^{k+1}-p^k\right)\right], \end{cases} \tag{1.18}$$

with p and q being the unknowns of the Hamiltonian formulation of the NLS equation (made from the decomposition $u = p + iq$). With $\tilde{g}_1^{k+1/2}$ and $\tilde{g}_2^{k+1/2}$ being second-order extrapolation of

$$g_1(p,q) = \frac{1}{\sqrt{\mathcal{E}_1[t]}}\frac{\delta \mathcal{E}_1[t]}{\delta q}, \quad g_2(p,q) = \frac{1}{\sqrt{\mathcal{E}_1[t]}}\frac{\delta \mathcal{E}_1[t]}{\delta p}.$$

Altogether, we expect our scheme to be second-order accurate in time and first-order in space. We prove that our scheme conserves a modified Hamiltonian energy unconditionally, and we are able to show that it also preserves the L^2 norm of the solution through time up to an error of order Δt^3.

From these conservation properties, and some simple inequalities derived from the scheme (and the conservation of the Hamiltonian), we present a standard convergence analysis and conduct an estimation of the error committed by the scheme

Theorem 9 (Error analysis) *Assuming that the solution of the SAV scheme with initial condition satisfying*

$$u^0 \in H^3(\Omega).$$

The discrete solution $\{P^{k+1}, Q^{k+1}\}$ *of the scheme satisfies the error estimate*

$$\left\|\nabla e_q^{k+1}\right\|_0^2 + \left\|\nabla e_p^{k+1}\right\|_0^2 + \left|e_r^{k+1}\right|^2 \leq C \exp\left(\left[1 - C\tau\right]^{-1} t^{k+1}\right)\left(\tau^4 + N^{-2}\right). \tag{1.19}$$

We verify this analytical result with numerical simulations of solitary waves. We also compare the orders of convergence and the errors committed on the Hamiltonian and the solution with well-known numerical methods for this equation: the Lie and Strang splitting methods. The SAV method can simulate solitons correctly, and we recover the correct orders of convergence. Using the solution given by the SAV scheme to compute numerically the value of the Hamiltonian (not the modified one), the Strang splitting method gives a better result. However, for the soliton test case, the modified Hamiltonian given by the SAV scheme is closer to the real Hamiltonian compared to the results given by both splitting techniques. This result holds even for long-time simulations.

We then conduct numerical experiments to observe how the solution of the SAV scheme behaves compared to the ones given by splitting techniques for particular choice of nonlinearities and for different regularities of the initial condition. For the latter, the SAV scheme presents a behavior close to the splitting techniques i.e. decreasing the regularity of the initial condition, the order of convergence in time decreases. However, for non-integer exponent on the nonlinearity $f(|u|^2) = \beta |u|^{\frac{2}{\gamma}}$, we observe that the SAV scheme preserves its second-order convergence in time whereas splitting techniques fail (see Figure 1.12).

Therefore, our study illustrates the favorable energy conservation of the SAV method compared to classical splitting schemes in certain applications.

1.6 Discussion and perspectives

Mathematical models can help to unveil crucial mechanisms behind biological phenomena. Our work aims to propose new mathematical models obtained from revisiting well-established models to add particular effects and analyze these equations that are often nonlinear and induce numerous difficulties. Then, to compare the models with biological experiments, numerical simulations are needed. However, to rely on the numerical results, the time and space discretizations must lead to a correct approximation of the continuous models. Therefore, structure-preserving methods must be the first choice. However, due to their complexity and computational cost, numerical schemes often used in Mathematical Biology are too many times oversimplifying methods. To solve this issue, we propose in this manuscript different methods to simulate efficiently and correctly mathematical models of living tissues.

1.6.1 Simulation of the relaxed-degenerate Cahn-Hilliard model and effect of the relaxation

To solve the computational cost issue of previous numerical methods for the Cahn-Hilliard equation with a biologically relevant choice of potential, we proposed a relaxation of the equation.

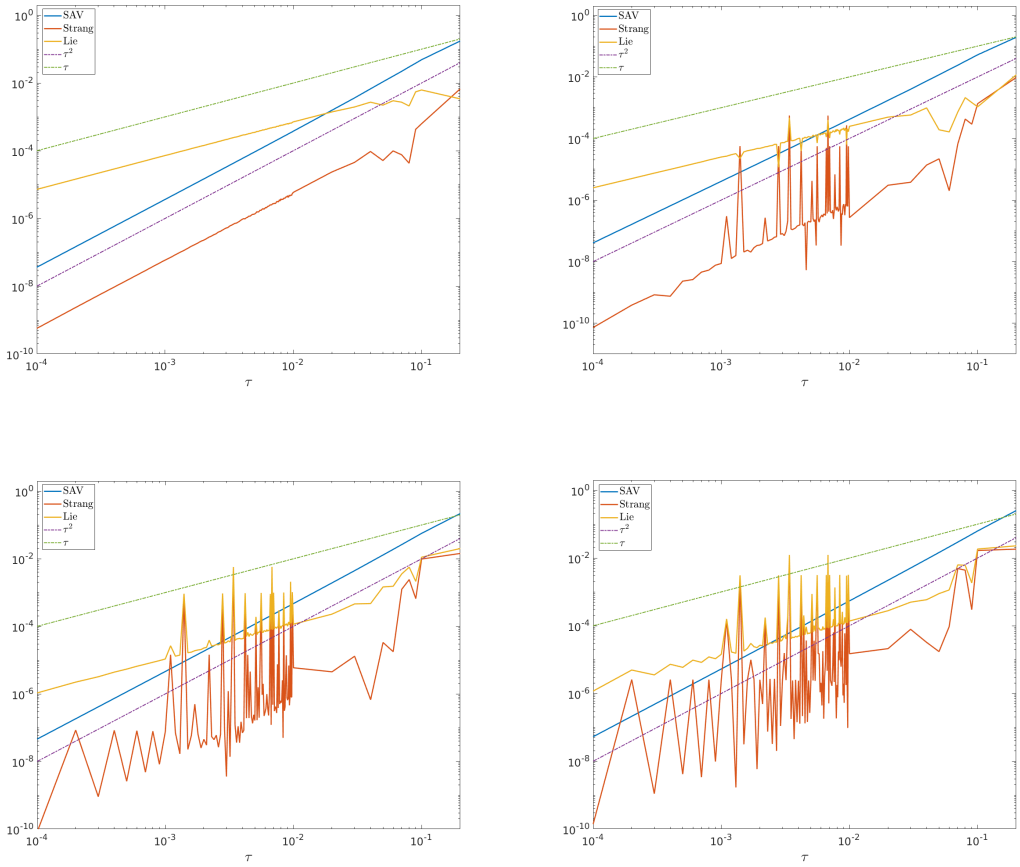


Figure 1.12 – Error on the Hamiltonian for various values of Δt for the nonlinear Schrödinger equation with different non-integer exponent ($\gamma = 2$ top-left, $\gamma = \frac{8}{3}$ top-right, $\gamma = 4$ bottom left and $\gamma = 8$ bottom-right). The dotted lines represent the order Δt (green) and Δt^2 (purple).

For this new model, we designed two structure-preserving numerical schemes. For the nonlinear scheme, we were able to show analytically that it retrieves the properties of the relaxed model. For the efficient linear scheme, the preservation of the structure of the model is observed during the numerical simulations. Still, we can not prove analytically that the discrete energy is dissipating. This property is essential to minimize the approximation error and is required to prove that the solution of the discrete scheme converges to one solution of the continuous model.

We aim to adapt the Scalar Auxiliary Variable method to our relaxed-degenerate Cahn-Hilliard model to solve this problem. Indeed, the RDCH model can be formulated as a gradient-flow of its associated energy, and the SAV method can thus be applied. We adapted the SAV method for the volume-filling Keller-Segel model, and showed that it allows to design an efficient linear scheme that preserves the dissipation of a modified energy. To make sure that the modified energy is close to the original one, we evaluated the error for a well-known nonlinear dispersive equation analytically and numerically. Our results for the Keller-Segel model and the Nonlinear Schrödinger equation show that the SAV method is robust, efficient, and if combined with an

upwind stabilization, enhances the spatial accuracy.

However, it remains unclear at the moment how to evaluate the error between the solutions of the RDCH model and of the DCH model due to the relaxation parameter σ. Indeed, understanding quantitatively the error introduced by the relaxation remains an open question. Furthermore, the comparison of our numerical scheme for the RDCH model with results of simulations of the original DCH model was only qualitative. To allow for a quantitative comparison, we need to identify a relevant quantity. We propose to compare the phase-ordering dynamics in two dimensions for simulations of the original model performed in [8] and the ones given in Chapter 3. We know from the literature that the coarsening domains follow a growth law of the form $L(t) \sim t^{\alpha}$. This law is estimated from the inverse of the first moment of the spherically averaged structure factor [41]

$$L(t) =< k >^{-1} = \frac{\int kS(k,t)\ \mathrm{d}k}{\int S(k,t)\ \mathrm{d}k},$$

with $S(k,t)$ is the spherically averaged time-dependent structure factor, and k being the wavevectors of the Fourier transform of the time-equal correlation function

$$C(r,t) =< n(x+r,t)n(x,t) > .$$

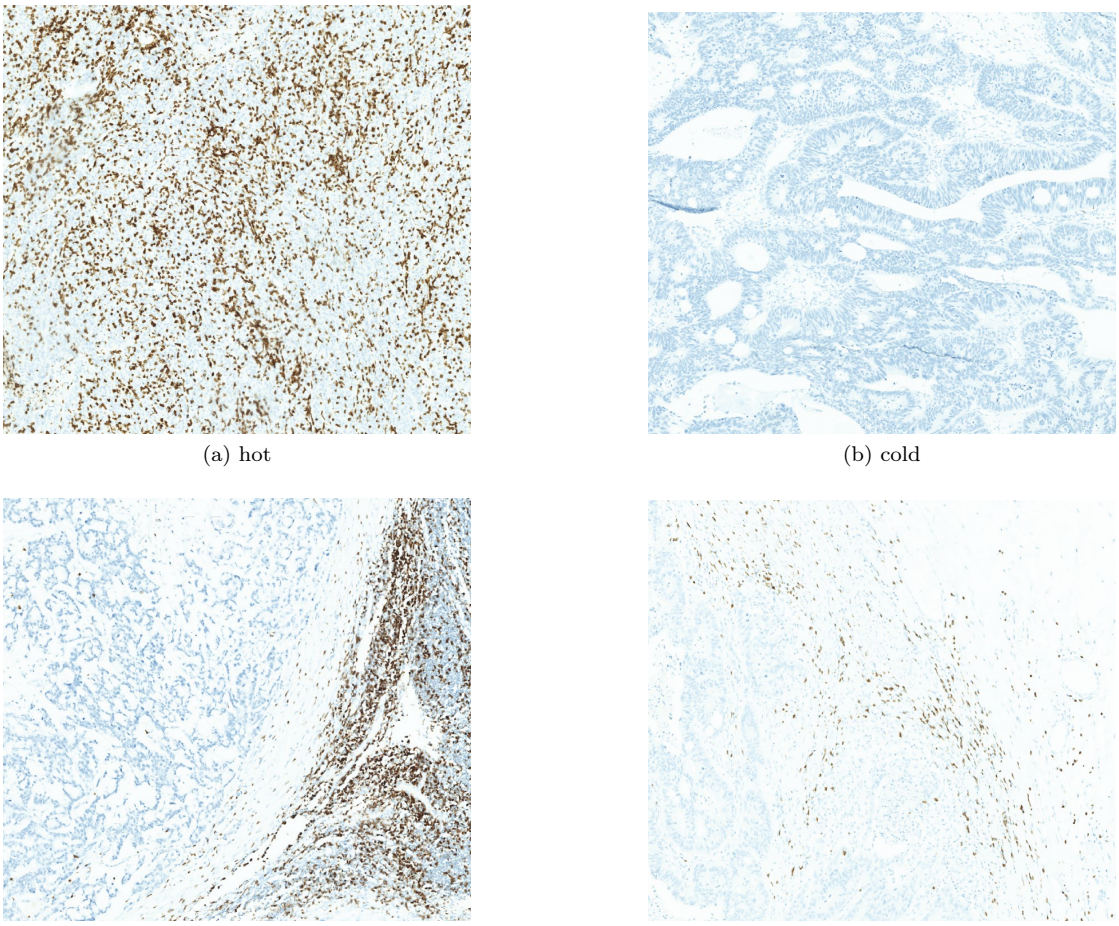
where $< \cdot >$ denotes ensemble averaging. For the DCH model with a single-well logarithmic potential, Agosti *et. al.* [8] indicated that the growth law is given by $L(t) \sim t^{0.3}$. Therefore, we doing the same computations, we will be able to compare quantitatively the numerical results for the DCH model and its relaxation.

1.6.2 Support a deeper understanding of key mechanisms in tumor progression

In this manuscript, we investigated the role of mechanical effects in the progression and organization of tumors. In particular, we proposed to give an explanation for two observed phenomena in the organization of tumor cells. On the one hand, to understand the shrinking of tumor cells due to a chemotherapeutic drug, we studied the assumption that tumor cells change their mechanical properties: from a solid to a semi-elastic body. Our work relies on the derivation and numerical simulations of a nonlinear volume-filling Keller-Segel model that takes into account the effect of the drug. On the other hand, to explain the formation of irregularities at the border of tumors during invasion processes, we proposed a mathematical model consistent with basic mechanics and thermodynamics. Our mathematical model is rather complicated but considers the effects of friction on the extracellular matrix, viscosity, attraction and repulsion between the cells, and proliferation. These two works proposed mathematical models that focus on physical effects as an explanation of the organization of tumor cells. In consequence, they are coarse approximations of the reality, and to get closer to the biological reality, it is necessary to take into account more effects.

Building upon the results of these two previous works, we are interested in investigating a particular scenario for in-vivo tumors that can help the development of a recent therapy. Indeed, recent researches in Medicine indicates that immunotherapy is a promising cure for malignant tumors. Different immunotherapy treatments exist: targeted antibodies, cancer vaccines, adoptive cell transfer, tumor-infecting viruses, checkpoint inhibitors, cytokines, and adjuvants. The response to this treatment depends on many factors, but one of the most important is the T lymphocytes' infiltration inside the tumor before the treatment. The different types of infiltration of tumors by lymphocytes is an indicator of the prognosis. Galon *et. al.* [96] proposed a classification in four categories (see Figure 1.13). The hot tumors are inflamed and infiltrated

with activated T cells even in the center of it. The category "altered-immunosuppressed" denotes tumors with a small amount of infiltrated T cells. Tumors that enter the "altered-excluded" category present different regions: their border is infiltrated by activated T cells while the center is deprived from lymphocytes. The last category is "cold" tumors and they are often correlated to a poor response to immunotherapy since no T cells are inside the tumor. However, very little is known about T cells' mechanisms and their different regulators to obtain the different observed patterns.



(a) hot

(b) cold

(c) altered-excluded

(d) altered-immunosuppressed

Figure 1.13 – Classification of tumors in function of T cells infiltration [96] (permission to reproduce the figures has been asked to the journal and we are waiting for their answer).)

Based on the development of multiphase Cahn-Hilliard system [38, 39, 37, 127, 97], and following the same approach as for the understanding of the irregularities during tumor invasion, we will propose a general three-phase Cahn-Hilliard model representing the interaction between tumor cells, immune cells, and the micro-environment. This model will consider the effect of cell-cell adhesion and repulsion through a physically relevant choice of potential. Based on the works of Boyer [38, 37], we will choose biologically relevant potentials for three-phase systems that remain physically consistent, i.e., no artificial emergence of a third phase within the interface

separating two phases. Indeed, our choice to add this effect of interaction between cells is motivated by biological observations indicating that adhesion between tumor cells exists, and adhesion between immune cells plays a central role in the recruiting and activation processes. Chemotaxis for the immune cells will also be taken into account since it is well-known that activated T cells release cytokines in the micro-environment that guide other T cells and play an activating or inhibitory effect on the recruiting and activation. Then, two source terms must be taken into account for the proliferation and death of the tumor cells, and the other for recruiting activated T cells. These two source terms must be carefully chosen to represent the different scenarios that can occur. For example, at a particular stage, the tumor cells are expected to escape immune-surveillance, and therefore inactivate T cells, having the consequence of decreasing their amount.

Our first calculations show that our mathematical model will be composed of two coupled Cahn-Hilliard equations: one for the evolution of the tumor cells and the other for the evolution of the immune cells (the micro-environment being determined from the two previous). To conduct numerical experiments, we will adapt our relaxation method, and structure-preserving numerical schemes. Other mechanical effects will be taken into account to understand their effects on the observed patterns of infiltration.

Part I

The Cahn-Hilliard equation for Biology

Chapter 2

Relaxation of the Cahn-Hilliard equation for Biology

Abstract

The degenerate Cahn-Hilliard equation is a standard model to describe living tissues. It takes into account cell populations undergoing short-range attraction and long-range repulsion effects. In this framework, we consider the usual Cahn-Hilliard equation with a singular single-well potential and degenerate mobility. These degeneracy and singularity induce numerous difficulties, in particular for its numerical simulation. To overcome these issues, we propose a relaxation system formed of two second order equations which can be solved with standard packages. This system is endowed with an energy and an entropy structure compatible with the limiting equation. Here, we study the theoretical properties of this system; global existence and convergence of the relaxed system to the degenerate Cahn-Hilliard equation. We also study the long-time asymptotics which interest relies on the numerous possible steady states with given mass.

This chapter is taken from Benoît Perthame, A. P., *Relaxation of the Cahn-Hilliard equation with singular single-well potential and degenerate mobility*, European Journal of applied mathematics (2020). Journal.

2.1 Introduction

The Degenerate Cahn-Hilliard equation (DCH in short) is a standard model, widely used in the mechanics of living tissues, [29, 203, 61, 8, 6, 92]. It is usual to set this problem in a smooth bounded domain $\Omega \subset \mathbb{R}^d$ with the zero flux boundary condition

$$\partial_t n = \nabla \cdot (b(n) \nabla (-\gamma \Delta n + \psi'(n))) \quad \text{in} \quad \Omega \times (0, +\infty), \tag{2.1}$$

$$\frac{\partial n}{\partial \nu} = b(n) \frac{\partial (-\gamma \Delta n + \psi'(n))}{\partial \nu} = 0 \qquad \text{on} \quad \partial\Omega \times (0, +\infty), \tag{2.2}$$

where ν is the outward normal vector to the boundary $\partial\Omega$ and $n = \frac{n_1}{n_1+n_2}$ represents the relative density or volume fraction of one of the two cell types.

Degeneracy of the coefficient $b(n)$ and singularity of the potential $\psi(n)$ make this problem particularly difficult to solve numerically and in particular, to preserve the apriori bound $0 \leqslant n < 1$. Motivated by the use of standard software for elliptic or parabolic equations, we propose to study

the following relaxed degenerate Cahn-Hilliard equation (RDHC in short)

$$\begin{cases} \partial_t n = \nabla \cdot \big(b(n)\nabla\left(\varphi + \psi_+'(n)\right)\big) & \text{in } \Omega \times (0,+\infty), \\ -\sigma\Delta\varphi + \varphi = -\gamma\Delta n + \psi_-'\big(n - \dfrac{\sigma}{\gamma}\varphi\big) & \text{in } \Omega \times (0,+\infty). \end{cases} \tag{2.3}$$

supplemented with zero-flux boundary conditions

$$\frac{\partial(\gamma n - \sigma\varphi)}{\partial\nu} = b(n)\frac{\partial\big(\varphi + \psi_+'(n)\big)}{\partial\nu} = 0 \qquad \text{on } \partial\Omega \times (0,+\infty). \tag{2.4}$$

Our purpose is to study existence for this system, to prove that as $\sigma \to 0$, the solution of RDCH system converges to the solution of the DCH equation and study the possible long term limits to steady states.

We make the following assumptions for the different inputs of the system (2.3). For the mechanics of living tissues, the usual assumption is that the potential ψ is concave degenerate near $n = 0$ (short-range attraction) and convex for n not too small (long-range repulsion). Additionally, a singularity at $n = 1$ is desired to represent saturation by one phase [46]. For these reasons, we call the potential *single-well logarithmic* and we decompose it in a convex and a concave part $\psi_\pm$

$$\psi(n) = \psi_+(n) + \psi_-(n), \qquad \pm\psi_\pm''(n) \geqslant 0, \qquad 0 \leqslant n < 1. \tag{2.5}$$

The singularity is contained in the convex part of the potential and we assume that

$$\psi_+ \in C^2\big([0,1)\big), \quad \psi_+'(1) = \infty, \tag{2.6}$$

and we extend the smooth concave part on $[0,1]$ to the full line with

$$\psi_- \in C^2(\mathbb{R}) \qquad \psi_-,\ \psi_-',\ \psi_-'' \quad \text{are bounded and } \frac{\sigma}{\gamma}||\psi_-''||_\infty < 1. \tag{2.7}$$

In practice, typical examples of potentials are, for some $n^* \in (0,1)$, see [63, 56]

$$\psi(n) = -(1-n^*)\ln(1-n) - \frac{n^3}{3} - (1-n^*)\frac{n^2}{2} - (1-n^*)n + k, \tag{2.8}$$

$$\psi(n) = \frac{1}{2}n\ln n + (1-n)\ln(1-n) - (n-\frac{1}{2})^2. \tag{2.9}$$

The potential (2.8) fulfills our assumptions and the convex/concave decomposition reads for $n \in [0,1)$

$$\psi_+(n) = -(1-n^*)\log(1-n) - \frac{n^3}{3}, \qquad \psi_-(n) = -(1-n^*)\frac{n^2}{2} - (1-n^*)n + k.$$

In this case ψ_+ is convex if $n^* \leq 0.7$. Potential (2.9) does not satisfy our assumptions because of the additional singularity at 0 (and thus is not treated here), however, it can also be decomposed as needed with

$$\psi_+(n) = \frac{1}{2}n\ln n + (1-n)\ln(1-n), \qquad \psi_-(n) = -(n-\frac{1}{2})^2.$$

To satisfy the assumptions (2.6) and (2.7), we need to extend the potential ψ_- to all $\mathbb{R}$ since the

above examples are defined for $n \in [0,1)$, which is an immediate task.

The potential (2.8) has been used to model the interaction between cancer cells from a glioblastoma multiforme and healthy cells by Agosti *et al.* [9] and promising results have been obtained. We also use the degeneracy assumption on $b \in C^1([0,1];\mathbb{R}^+)$,

$$b(0) = b(1) = 0, \qquad b(n) > 0 \text{ for } 0 < n < 1. \tag{2.10}$$

The typical expression in the applications we have in mind is $b(n) = n(1-n)^2$. Consequently, when considered as transport equations, both (2.1) and (2.3) impose formally the property that $0 \leqslant n \leqslant 1$. However, we need an additional technical assumption, namely that there is some cancellation at 1 such that

$$b(\cdot)\psi''(\cdot) \in C([0,1];\mathbb{R}). \tag{2.11}$$

We implicitly assume (2.5)–(2.11) in this paper. Also, we always impose an initial condition satisfying

$$n^0 \in H^1(\Omega), \qquad 0 \leq n^0 < 1 \quad \text{a.e. in } \Omega. \tag{2.12}$$

The assumption $n^0 \in [0,1)$ is consistent with the degeneracy of mobility at 0 which allows solutions to vanish on open sets. But the singularity of the potential at 1 and the energy bound make that $n = 1$ cannot be achieved except of a negligible set. Thanks to the boundary condition (2.2), the system conserves the initial mass

$$\int_\Omega n(x,t)dx = \int_\Omega n^0(x)dx =: M, \quad \forall t \geq 0.$$

We denote the flux associated with the RDCH system by

$$J_\sigma(n,\varphi) := -b(n)\nabla\left(\varphi + \psi_+'(n)\right). \tag{2.13}$$

The system (2.3) comes with energy and entropy structures, namely, the energy is defined as

$$\mathcal{E}_\sigma[n_\sigma] = \int_\Omega \left[\psi_+(n_\sigma) + \frac{\gamma}{2}|\nabla(n_\sigma - \frac{\sigma}{\gamma}\varphi_\sigma)|^2 + \frac{\sigma}{2\gamma}|\varphi_\sigma|^2 + \psi_-(n_\sigma - \frac{\sigma}{\gamma}\varphi_\sigma)\right]. \tag{2.14}$$

The energy is bounded from below thanks to the assumptions above and satisfies

$$\frac{d}{dt}\mathcal{E}_\sigma[n_\sigma(t)] = -\int_\Omega b(n_\sigma)\left|\nabla(\varphi_\sigma + \psi_+'(n_\sigma))\right|^2 \leqslant 0. \tag{2.15}$$

For the entropy, we set for $0 < n < 1$ the singular function

$$\phi''(n) = \frac{1}{b(n)}, \qquad \Phi[n] = \int_\Omega \phi\big(n(x)\big)dx. \tag{2.16}$$

The entropy functional behaves as follows in the case $b(n) = n(1-n)^2$

$$\phi(n) = n\log(n),\ n \approx 0^+, \qquad \phi(n) = -\log(1-n),\ n \approx 1^-.$$

The relation holds

$$\begin{aligned}\frac{d\Phi[n_\sigma(t)]}{dt} = -\int_\Omega \gamma\left|\Delta\left(n_\sigma - \frac{\sigma}{\gamma}\varphi_\sigma\right)\right|^2 + \frac{\sigma}{\gamma}|\nabla\varphi_\sigma|^2 + \psi_-''(n_\sigma - \frac{\sigma}{\gamma}\varphi_\sigma)\left|\nabla(n_\sigma - \frac{\sigma}{\gamma}\varphi_\sigma)\right|^2 \\ + \psi_+''(n_\sigma)|\nabla n_\sigma|^2.\end{aligned} \tag{2.17}$$

Notice that entropy equality does not provide us with a direct a priori estimate because of the term ψ_-'' can be negative. Therefore we have to combine it with the energy dissipation to write

$$\Phi[n_\sigma(T)] + \int_{\Omega_T} \left[\gamma \left| \Delta \left(n_\sigma - \frac{\sigma}{\gamma} \varphi_\sigma \right) \right|^2 + \frac{\sigma}{\gamma} |\nabla \varphi_\sigma|^2 + \psi_+''(n_\sigma) |\nabla n_\sigma|^2 \right] \leqslant \Phi[n^0] + \frac{2T}{\gamma} \|\psi_-''\|_\infty \, \mathcal{E}_\sigma[n^0].$$

The first use of the Cahn-Hilliard equation is to model the spinodal decomposition occurring in binary materials during a sudden cooling [48, 47]. The bilaplacian $-\gamma\Delta^2 n$ is used to represent surface tension and the parameter γ is the square of the width of the diffuse interface between the two phases. In both equations (2.1) and (2.3), $n = n(x,t)$ is a relative quantity: for our biological application this represents a relative cell density as derived from phase-field models [46] and for this reason the property $n \in [0,1)$ is relevant. The biological explanation of the fact that 1 is excluded from the interval of definition of n is due to the observation that cells tend to not form aggregates that are too dense. For instance, the two phases can be the relative density of cancer cells and the other component represents the extracellular matrix, liquid, and other cells. This binary mixture tends to form aggregates in which the density of one component of the binary mixture is larger than the other component. The interest of the Cahn-Hilliard equation stems from solutions that reproduce the formation of such clusters of cells *in vivo* or on dishes. Several variants are also used. A Cahn-Hilliard-Hele-Shaw model is proposed by Lowengrub *et al* [142] to describe the avascular, vascular and metastatic stages of solid tumor growth. They proved the existence and uniqueness of a strong solution globally for $d \leq 2$ and locally for $d = 3$ as well as the long term convergence to steady-state. The case with a singular potential is treated in [102]. Variants can include the coupling with fluid equations and chemotaxis, see for instance [71] and the references therein.

The analysis of the long-time behavior of the solution of the Cahn-Hilliard equation has also attracted much attention since the seminal paper [35]. A precise description of the ω-limit set has been obtained in one dimension for the case of smooth polynomial potential and constant mobility in [183]. In this work, the effect of the different parameters of the model such as the initial mass, the width of the diffuse interface are investigated. In fact, the authors show that when γ is large, the solution converges to a constant as $t \to \infty$. The same happens when the initial mass is large. However when γ is positive and small enough, the system admits nontrivial steady-states. For logarithmic potentials and constant mobility, Abels and Wilke [2] prove that solutions converge to a steady-state as time goes to infinity using the Lojasiewicz–Simon inequality. Other works have been made on the long term behavior of the solutions of some Cahn-Hilliard models including a source term [59], with dynamic boundary conditions [101], coupled with the Navier-Stokes equation [95], for non-local interactions and a reaction term [121].

Many difficulties, both analytical and numerical, arise in the context of Cahn-Hilliard equation and its variants. Because of the bilaplacian term, most of the numerical methods require to change the equation (2.1) into a system of two coupled equations

$$\begin{cases} \partial_t n = \nabla \cdot (b(n) \nabla v) \,, \\ \quad v = -\gamma \Delta n + \psi'(n). \end{cases} \tag{2.18}$$

This system of equations has been analyzed in the case where the mobility is degenerate and the potential is a logarithmic double-well functional by Elliott and Garcke [76]. They establish the existence of weak solutions of this system. Agosti *et al* [8] establish the existence of weak solutions

when ψ is a single-well logarithmic potential which is more relevant for biological applications (see [46]). They also prove that this system preserves the positivity of the cell density and the weak solutions belong to

$$n \in L^\infty(0,T;H^1(\Omega)) \cap L^2(0,T;H^2(\Omega)) \cap H^1(0,T;(H^1(\Omega))'), \quad J \in L^2((0,T)\times\Omega,\mathbb{R}^d) \quad \forall T>0,$$

The Cahn-Hilliard equation can be seen as an approximation of the famous microscopic model in [99, 100]. With our notations, it reads

$$\partial_t n = \nabla \cdot \left[b(n) \nabla \big(K_\sigma \star n + \psi'(n) \big) \right],$$

with a symmetric smooth kernel $K_\sigma \underset{\sigma\to 0}{\longrightarrow} \Delta\delta$. The convergence to the DCH equation has been answered recently in [67] in the case of periodic boundary conditions. Although, very similar in its form, our relaxation model undergoes different a priori estimates which allow us to study differently the limit $\sigma \to 0$ for (2.3).

For a full review about the mathematical analysis of the Cahn-Hilliard equation and its variants, we refer the reader to the recent book of Miranville [147].

Numerical simulations of the DCH system have been also performed in the context of double-well potentials in [77, 28]. To keep the energy inequality is a major concern in numerical methods and the survey paper by Shen *et al* [180] presents a general method applied to the present context.

Numerics is also our motivation to propose a relaxation of equation (2.1) in a form close to the writing (2.18). We recover the system (2.3) by introducing a new potential φ and a regularizing equation which defines v through $\nabla\varphi$. We use the decomposition (2.5) of the potential to keep the convex and stable part in the main equation for n, rejecting the concave and unstable part in the regularized equation. The relaxation parameter is σ and we need to verify that, in the limit $\sigma \to 0$, we recover the original DCH equation (2.1). This is the main purpose of the present paper.

As a first step towards the existence of solutions of (2.3), in section 2.2, we introduce a regularized problem which is not anymore degenerate and we prove the existence of weak solutions for this regularized-relaxed Cahn-Hilliard system. We show energy and entropy estimates from which we obtain a priori estimates which are used later on. In section 2.3, we pass to the limit in the regularization parameter ϵ and show the existence of weak solutions of the RDCH system. Then, in section 2.4, we prove the convergence as $\sigma \to 0$ to the full DCH model. Section 2.5 is dedicated to the study of the long term convergence of the solutions to steady-states. We end the paper with some conclusions and perspectives.

2.2 The regularized problem

To prove that the system (2.3), admits solutions and to precise the functional spaces, we first define a regularized problem. Then we prove the existence of solutions and estimates based on energy and entropy relations.

2.2.1 Regularization procedure

We consider a small positive parameter $0 < \epsilon \ll 1$ and define the regularized mobility

$$B_\epsilon(n) = \begin{cases} b(1-\epsilon) & \text{for } n \geq 1-\epsilon, \\ b(\epsilon) & \text{for } n \leq \epsilon, \\ b(n) & \text{otherwise.} \end{cases} \tag{2.19}$$

Then, there are two positive constants b_1 and B_1, such that

$$b_1 < B_\epsilon(n) < B_1, \quad \forall n \in \mathbb{R}. \tag{2.20}$$

Thus, the regularized mobility satisfies

$$B_\epsilon \in C(\mathbb{R}, \mathbb{R}^+). \tag{2.21}$$

To define a regular potential, we smooth out the singularity located at $n = 1$ which only occurs in ψ_+, see (2.6)–(2.7), and preserve the assumption (2.11) by setting

$$\psi''_{+,\epsilon}(n) = \begin{cases} \psi''_+(1-\epsilon) & \text{for } n \geq 1-\epsilon, \\ \psi''_+(\epsilon) & \text{for } n \leq \epsilon, \\ \psi''_+(n) & \text{otherwise.} \end{cases} \tag{2.22}$$

It is useful to notice that, for some positive constants D_1 independent of $0 < \epsilon \leqslant \epsilon_0$ and D_ϵ, we have

$$\psi_{+,\epsilon}(n) \in C^2(\mathbb{R}, \mathbb{R}) \quad \psi_{+,\epsilon}(n) \geqslant -D_1, \quad |\psi'_\epsilon(n)| \leq D_\epsilon(1 + |n|), \quad \forall n \in \mathbb{R}. \tag{2.23}$$

See also [8] for details about the extensions needed for the potential (2.8).

We can now define the regularized problem

$$\begin{cases} \partial_t n_{\sigma,\epsilon} = \nabla \cdot \left[B_\epsilon(n_{\sigma,\epsilon}) \nabla (\varphi_{\sigma,\epsilon} + \psi'_{+,\epsilon}(n_{\sigma,\epsilon})) \right], \\ -\sigma \Delta \varphi_{\sigma,\epsilon} + \varphi_{\sigma,\epsilon} = -\gamma \Delta n_{\sigma,\epsilon} + \psi'_-(n_{\sigma,\epsilon} - \dfrac{\sigma}{\gamma} \varphi_{\sigma,\epsilon}), \end{cases} \tag{2.24}$$

with zero-flux boundary conditions

$$\frac{\partial (n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon})}{\partial \nu} = \frac{\partial \left(\varphi_{\sigma,\epsilon} + \psi'_{+,\epsilon}(n_{\sigma,\epsilon})\right)}{\partial \nu} = 0 \qquad \text{on } \partial\Omega \times (0, +\infty). \tag{2.25}$$

It is convenient to define the flux of the regularized system as

$$J_{\sigma,\epsilon} = -B_\epsilon(n_{\sigma,\epsilon}) \nabla \left(\varphi_{\sigma,\epsilon} + \psi'_{+,\epsilon}(n_{\sigma,\epsilon}) \right).$$

2.2.2 Existence for the regularized problem

We can now state the existence theorem for the regularized problem (2.24).

Theorem 10 (Existence for $\epsilon > 0$) *Assuming $n^0 \in H^1(\Omega)$, there exists a pair of functions $(n_{\sigma,\epsilon}, \varphi_{\sigma,\epsilon})$ such that for all $T > 0$,*

$$\begin{aligned} n_{\sigma,\epsilon} &\in L^2(0,T; H^1(\Omega)), & \partial_t n_{\sigma,\epsilon} &\in L^2(0,T; (H^1(\Omega))'), \\ \varphi_{\sigma,\epsilon} &\in L^2(0,T; H^1(\Omega)), & & \\ n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon} &\in L^2(0,T; H^2(\Omega)), & \partial_t \left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon} \right) &\in L^2(0,T; (H^1(\Omega))'), \end{aligned}$$

which satisfies the regularized-relaxed degenerate Cahn-Hilliard equation (2.24), (2.25) in the

following weak sense: for all test function $\chi \in L^2(0,T;H^1(\Omega))$, it holds

$$\begin{aligned} \int_0^T < \chi, \partial_t n_{\sigma,\epsilon} > &= \int_{\Omega_T} B_\epsilon(n_{\sigma,\epsilon}) \nabla \left(\varphi_{\sigma,\epsilon} + \psi'_{+,\epsilon}(n_{\sigma,\epsilon}) \right) \nabla \chi, \\ \sigma \int_{\Omega_T} \nabla \varphi_{\sigma,\epsilon} \nabla \chi + \int_{\Omega_T} \varphi_{\sigma,\epsilon} \chi &= \gamma \int_{\Omega_T} \nabla n_{\sigma,\epsilon} \nabla \chi + \int_{\Omega_T} \psi'_-(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma} \varphi_{\sigma,\epsilon}) \chi. \end{aligned} \tag{2.26}$$

Proof. We adapt the proof of the theorem 2 in [76] where the authors prove the existence of solutions of the Cahn-Hilliard system with positive mobilities. Since the regularized mobility here is positive due to (2.20), we can apply the same theorem. The proof of existence follows the following different stages

Step 1. Galerkin approximation. Firstly, we make an approximation of the regularized problem (2.24). We define the family of eigenfunctions $\{\phi_i\}_{i\in\mathbb{N}}$ of the Laplace operator subjected to zero Neumann boundary conditions.

$$-\Delta \phi_i = \lambda_i \phi_i \text{ in } \Omega \quad \text{ with } \quad \nabla \phi_i \cdot \nu = 0 \text{ on } \partial\Omega.$$

The family $\{\phi_i\}_{i\in\mathbb{N}}$ form an orthogonal basis of both $H^1(\Omega)$ and $L^2(\Omega)$ and we normalize them, i.e. $(\phi_i, \phi_j)_{L^2(\Omega)} = \delta_{ij}$ to obtain an orthonormal basis. We assume that the first eigenvalue is $\lambda_1 = 0$ (which does not introduce a lack of generality).

We consider the following discretization of (2.24)

$$n^N(t,x) = \sum_{i=1}^N c_i^N(t)\phi_i(x), \quad \varphi^N(t,x) = \sum_{i=1}^N d_i^N(t)\phi_i(x), \tag{2.27}$$

$$\int_\Omega \partial_t n^N \phi_j = -\int_\Omega B_\epsilon(n^N) \nabla \left(\varphi^N + \Pi^N \left(\psi'_{+,\epsilon}(n^N) \right) \right) \nabla \phi_j, \quad \text{for } j = 1, ..., N, \tag{2.28}$$

$$\int_\Omega \varphi^N \phi_j = \gamma \int_\Omega \nabla \left(n^N - \frac{\sigma}{\gamma} \varphi^N \right) \nabla \phi_j + \int_\Omega \psi'_-(n^N - \frac{\sigma}{\gamma}\varphi^N)\phi_j, \quad \text{for } j = 1, ..., N, \tag{2.29}$$

$$n^N(0,x) = \sum_{i=1}^N (n_0, \phi_i)_{L^2(\Omega)} \phi_i. \tag{2.30}$$

We have used the L^2 projection $\Pi^N : L^2(\Omega) \to V$, where $V = \text{span}\{\phi_1, ..., \phi_N\}$. This gives the following initial value problem for a system of ordinary differential equations, for all $j = 1, ..., N$,

$$\partial_t c_j^N = -\int_\Omega B_\epsilon(\sum_{i=1}^N c_i^N \phi_i) \nabla \left(\varphi^N + \Pi^N \left(\psi'_{+,\epsilon}(\sum_{i=1}^N c_i^N \phi_i) \right) \right) \nabla \phi_j, \tag{2.31}$$

$$d_j^N = \gamma \lambda_j c_j^N - \sigma \lambda_j d_j^N + \int_\Omega \psi'_-(\sum_{k=1}^N (c_k^N - \frac{\sigma}{\gamma} d_k^N)\phi_k)\phi_j, \tag{2.32}$$

$$c_j^N(0) = (n_0, \phi_j)_{L^2(\Omega)}. \tag{2.33}$$

Since the right-hand side of equation (2.31) depends continuously on the coefficients c_j^N, the initial value problem has a local solution.

Step 2. Inequalities and convergences. Multiplying equation (2.31), by $\phi_i\big(\varphi^N + \psi'_+(n^N)\big)$, then

summing over i and integrating over the domain leads to

$$\frac{d}{dt}\int_\Omega \psi_{+,\epsilon}(n^N) + \int_\Omega \partial_t(n^N)\varphi^N = \int_\Omega \sum_i (\varphi^N + \psi'_{+,\epsilon}(n^N))\phi_i \int_\Omega \nabla\phi_i \left(B_\epsilon(n^N)\nabla\left(\varphi^N + \Pi^N\left(\psi'_{+,\epsilon}(n^N)\right)\right)\right) \mathrm{d}y\ \mathrm{d}x. \tag{2.34}$$

Let us focus on the left-hand side with

$$\int_\Omega \partial_t(n^N)\varphi^N = \int_\Omega \partial_t(n^N - \frac{\sigma}{\gamma}\varphi^N)\varphi^N + \frac{1}{2}\frac{\sigma}{\gamma}\frac{d}{dt}\int_\Omega |\varphi^N|^2.$$

Then, using the equation (2.29), we have that

$$\int_\Omega \partial_t(n^N - \frac{\sigma}{\gamma}\varphi^N)\varphi^N = \frac{\gamma}{2}\frac{d}{dt}\int_\Omega |\nabla(n^N - \frac{\sigma}{\gamma}\varphi^N)|^2 + \frac{d}{dt}\int_\Omega \psi_-(n^N - \frac{\sigma}{\gamma}\varphi^N).$$

The right-hand side of equation (2.34) gives

$$\begin{aligned} &-\int_\Omega \sum_i (\varphi^N + \psi'_{+,\epsilon}(n^N))\phi_i \int_\Omega \nabla\phi_i \left(B_\epsilon(n^N)\nabla\left(\varphi^N + \Pi^N\left(\psi'_{+,\epsilon}(n^N)\right)\right)\right) \mathrm{d}y\ \mathrm{d}x \\ &\qquad = -\int_\Omega B_\epsilon(n^N)\left|\nabla\left(\varphi^N + \Pi^N\left(\psi'_{+,\epsilon}(n^N)\right)\right)\right|^2. \end{aligned}$$

Altogether, we obtain

$$\frac{d}{dt}E(t) + \int_\Omega B_\epsilon(n^N)\left|\nabla\left(\varphi^N + \Pi^N\left(\psi'_{+,\epsilon}(n^N)\right)\right)\right|^2 \le 0, \tag{2.35}$$

where

$$E(t) = \int_\Omega \psi_{+,\epsilon}(n^N) + \frac{\gamma}{2}\int_\Omega |\nabla(n^N - \frac{\sigma}{\gamma}\varphi^N)|^2 + \frac{1}{2}\frac{\sigma}{\gamma}\int_\Omega |\varphi^N|^2 + \int_\Omega \psi_-(n^N - \frac{\sigma}{\gamma}\varphi^N).$$

Next, to prove the compactness in space of ∇n^N, we write

$$\begin{aligned} \min_{n^N}\left(\frac{1+\frac{\sigma}{\gamma}\psi''_{+,\epsilon}}{\psi''_{+,\epsilon}}\right)^2 \int_\Omega \left|\nabla\psi'_{+,\epsilon}(n^N)\right|^2 &\le \int_\Omega \left(\frac{1+\frac{\sigma}{\gamma}\psi''_{+,\epsilon}}{\psi''_{+,\epsilon}}\right)^2 \left|\nabla\psi'_{+,\epsilon}(n^N)\right|^2 \\ &\le \int_\Omega \left|\nabla\left(n^N + \frac{\sigma}{\gamma}\psi'_{+,\epsilon}(n^N)\right)\right|^2. \end{aligned}$$

Therefore, for some $\theta > 0$, we have

$$\begin{aligned} \left(\left(\frac{\sigma}{\gamma}\right)^2 + \theta\right)\int_\Omega \left|\nabla\psi'_{+,\epsilon}(n^N)\right|^2 \le \int_\Omega \Big|\nabla\left(n^N - \frac{\sigma}{\gamma}\varphi^N\right) + \frac{\sigma}{\gamma}\nabla\left(\varphi^N + \Pi^N\left(\psi'_{+,\epsilon}(n^N)\right)\right) \\ + \frac{\sigma}{\gamma}\nabla\left(\psi'_{+,\epsilon}(n^N) - \Pi^N\left(\psi'_{+,\epsilon}(n^N)\right)\right)\Big|^2. \end{aligned}$$

Finally, we obtain

$$\left(\left(\frac{\sigma}{\gamma}\right)^2+\theta\right)\int_\Omega\left|\nabla\psi'_{+,\epsilon}(n^N)\right|^2\leq C(T)+\left(\frac{\sigma}{\gamma}\right)^2\int_\Omega\left|\nabla\left(\psi'_{+,\epsilon}(n^N)-\Pi^N\left(\psi'_{+,\epsilon}(n^N)\right)\right)\right|^2$$
$$\leq C(T)+\left(\frac{\sigma}{\gamma}\right)^2\int_\Omega\left|\nabla\psi'_{+,\epsilon}(n^N)\right|^2,$$

and we proved that

$$\theta\int_\Omega\left|\nabla\psi'_{+,\epsilon}(n^N)\right|^2\leq C(T).$$

Therefore, we can obtain from the previous inequalities the following

$$\frac{\gamma}{2}\int_\Omega|\nabla(n^N-\frac{\sigma}{\gamma}\varphi^N)|^2\leq C, \tag{2.36}$$
$$\frac{\sigma}{2\gamma}\int_\Omega|\varphi^N|^2\leq C, \tag{2.37}$$
$$\int_{\Omega_T}B_\epsilon(n^N)\left|\nabla\left(\varphi^N+\Pi^N\left(\psi'_{+,\epsilon}(n^N)\right)\right)\right|^2\leq C, \tag{2.38}$$
$$\theta\min_{r\in\mathbb{R}}\left(\psi''_{+,\epsilon}(r)\right)\int_\Omega|\nabla n^N|^2\leq C(T), \tag{2.39}$$

which hold for positive values of γ,σ,θ and also for all finite time $T\geq 0$. Therefore, from these inequalities we can extract subsequences of (n^N,φ^N) such that the following convergences hold for any time $T\geq 0$ and small positive values of γ,σ.

Taking $j=1$ in (2.28), gives the results that $\frac{d}{dt}\int n^N=0$. Then, using the inequality (2.39) and the Poincaré-Wirtinger inequality, we obtain

$$n^N\rightharpoonup n_{\sigma,\epsilon}\text{ weakly in }L^2(0,T;H^1(\Omega)). \tag{2.40}$$

This result, in turn, implies that the coefficients c_j^N are bounded and a global solution to (2.31)–(2.33) exists. Choosing $j=1$ in (2.29) gives

$$\int_\Omega\varphi^N=\int_\Omega\psi_-\left(n^N-\frac{\sigma}{\gamma}n^N\right),$$

and combining (2.36), (2.40) and the Poincaré-Wirtinger inequality gives

$$\varphi^N\rightharpoonup\varphi_{\sigma,\epsilon}\text{ weakly in }L^2(0,T;H^1(\Omega)). \tag{2.41}$$

We also obtain from (2.40) and (2.41)

$$n^N-\frac{\sigma}{\gamma}\varphi^N\rightharpoonup n_{\sigma,\epsilon}-\frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\text{ weakly in }L^2(0,T;H^1(\Omega)). \tag{2.42}$$

From the previous convergence, we conclude that $\varphi_{\sigma,\epsilon}\in L^2(0,T;H^1(\Omega))$, therefore, using elliptic regularity we know that

$$n_{\sigma,\epsilon}-\frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\in L^2(0,T;H^2(\Omega)). \tag{2.43}$$

To be able to prove some strong convergence in $L^2(0,T;L^2(\Omega))$ of n^N, we need an information

about the temporal derivative $\partial_t n^N$. From the first equation of the system, we have for all test functions $\phi \in L^2(0,T;H^1(\Omega))$

$$\begin{aligned}\left|\int_{\Omega_T} \partial_t n^N \phi\right| &= \left|\int_{\Omega_T} \partial_t n^N \Pi_N \phi\right| \\ &= \left|\int_{\Omega_T} b(n^N)\nabla\left(\varphi^N + \Pi^N\left(\psi'_{+,\epsilon}(n^N)\right)\right)\nabla\Pi_N\phi\right| \\ &\leq \left(B_1\int_{\Omega_T} B_\epsilon(n^N)\left|\nabla\left(\varphi^N + \Pi^N\left(\psi'_{+,\epsilon}(n^N)\right)\right)\right|^2\right)^{\frac{1}{2}}\left(\int_{\Omega_T}|\nabla\Pi_N\phi|^2\right)^{\frac{1}{2}}.\end{aligned} \tag{2.44}$$

Using (2.38), we obtain

$$\left|\int_{\Omega_T} \partial_t n^N \phi\right| \leq C\left(\int_{\Omega_T}|\nabla\Pi_N\phi|^2\right)^{\frac{1}{2}}. \tag{2.45}$$

Thus we can extract a subsequence such that

$$\partial_t n^N \rightharpoonup \partial_t n_{\sigma,\epsilon} \text{ weakly in } L^2(0,T;(H^1(\Omega))'). \tag{2.46}$$

From (2.40) and (2.46) and using the Lions-Aubin Lemma, we obtain the strong convergence

$$n^N \to n_{\sigma,\epsilon} \text{ strongly in } L^2(0,T;L^2(\Omega)). \tag{2.47}$$

Next, we need to prove the strong convergence of $n^N - \frac{\sigma}{\gamma}\varphi^N$ in $L^2(0,T;H^1(\Omega))$. In order to do that we must bound the $L^2(0,T;(H^1(\Omega))')$ norm of its time derivative. Starting from the equation (2.32), multiplying it by $-\frac{\sigma}{\gamma}$, adding c_j^N and calculating its time derivative, we obtain

$$\frac{d}{dt}\left(c_j^N - \frac{\sigma}{\gamma}d_j^N\right) = \frac{d}{dt}c_j^N - \sigma\lambda_j\frac{d}{dt}\left(c_j^N - \frac{\sigma}{\gamma}d_j^N\right) - \frac{\sigma}{\gamma}\frac{d}{dt}\int_\Omega \psi'_-\left(n^N - \frac{\sigma}{\gamma}\varphi^N\right)\phi_j.$$

Multiplying the previous equation by $\phi_j\partial_t\left(n^N - \frac{\sigma}{\gamma}\varphi^N\right)$, summing over j and integrating over Ω, we obtain

$$\begin{aligned}\int_\Omega\left(\partial_t\left(n^N - \frac{\sigma}{\gamma}\varphi^N\right)\right)^2 + \sigma\int_\Omega\left|\nabla\left(\partial_t\left(n^N - \frac{\sigma}{\gamma}\varphi^N\right)\right)\right|^2 &= \int_\Omega \partial_t n^N \partial_t\left(n^N - \frac{\sigma}{\gamma}\varphi^N\right) \\ &- \sum_j\int_\Omega \phi_j\partial_t\left(n^N - \frac{\sigma}{\gamma}\varphi^N\right)\frac{\sigma}{\gamma}\frac{d}{dt}\int_\Omega\psi'_-\left(n^N - \frac{\sigma}{\gamma}\varphi^N\right)\phi_j\,dx\,dy.\end{aligned}$$

Let us define $U^N = \partial_t\left(n^N - \frac{\sigma}{\gamma}\varphi^N\right)$ and rewrite the previous equation

$$\sigma\int_\Omega|\nabla U^N|^2 + \int_\Omega|U^N|^2 = \int_\Omega \partial_t n^N U^N - \frac{\sigma}{\gamma}\int_\Omega|U^N|^2\psi''_-(n^N - \frac{\sigma}{\gamma}\varphi^N).$$

From the Cauchy-Schwarz inequality, we obtain

$$0 \leq ||\nabla U^N||^2_{L^2(\Omega)} + \left(1 - \frac{\sigma}{\gamma}||\psi''_-||_\infty\right)||U^N||^2_{L^2(\Omega)} \leq ||\partial_t n^N||_{L^2(\Omega)}||U^N||_{L^2(\Omega)}. \tag{2.48}$$

Finally, from the (2.7) we obtain that

$$||U^N||_{L^2(0,T;(H^1(\Omega))')} \leq C.$$

Therefore, we can extract a subsequence such that

$$\partial_t \left(n^N - \frac{\sigma}{\gamma}\varphi^N \right) \rightharpoonup \partial_t \left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon} \right) \text{ weakly in } L^2(0,T;(H^1(\Omega))'). \tag{2.49}$$

Using (2.49) and (2.43) and the Lions-Aubin lemma we obtain the following strong convergence

$$n^N - \frac{\sigma}{\gamma}\varphi^N \to n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon} \text{ strongly in } L^2(0,T;H^1(\Omega)). \tag{2.50}$$

Step 3. Limiting equation. The main difficulty to pass to the limit in the equation (2.29) relies mainly on the convergence of the term $\int_\Omega \psi'_-(n^N - \frac{\sigma}{\gamma}\varphi^N)\phi_j$ which is solved using the strong convergence (2.50) and the properties (2.7). Therefore, we obtain

$$\psi'_-(n^N - \frac{\sigma}{\gamma}\varphi^N) \to \psi'_-(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}) \quad \text{a.e. in } \Omega_T. \tag{2.51}$$

Then combining the convergences (2.42), (2.41), (2.51) and the Lebesgue dominated convergence theorem, we pass to the limit in the equation (2.29). We can also pass to the limit in the first equation (2.28) by the standard manner (see [134]), using the strong convergence (2.47), the properties of the mobility (2.21) and the potential (2.23). Altogether, we obtain the limiting system (2.26).

2.2.3 Energy, entropy and a priori estimates

The relaxed and regularized system (2.24) comes with an energy and an entropy. These provide us with estimates which are useful to prove the existence of global weak solutions of (2.24) and their convergence to the weak solutions of the original DHC equation or to the RDHC as ϵ and/or $\sigma \to 0$.

Being given a solution $(n_{\sigma,\epsilon}, \varphi_{\sigma,\epsilon})$ satisfying Theorem 10, we define the energy associated with the regularized potential $\psi_{+,\epsilon}$ and relaxed system as

$$\mathcal{E}_{\sigma,\epsilon}[n_{\sigma,\epsilon}] = \int_\Omega \left[\psi_{+,\epsilon}(n_{\sigma,\epsilon}) + \frac{\gamma}{2}|\nabla(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon})|^2 + \frac{\sigma}{2\gamma}|\varphi_{\sigma,\epsilon}|^2 + \psi_-(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}) \right], \tag{2.52}$$

where $\varphi_{\sigma,\epsilon}$ is obtained from $n_{\sigma,\epsilon}$ by solving the elliptic equation in (2.24). Notice that $\mathcal{E}_{\sigma,\epsilon}[n_{\sigma,\epsilon}]$ is lower bounded, uniformly in ϵ and σ, thanks to the assumptions on ψ_- in (2.7) and the construction of $\psi_{\epsilon,+}$ in (2.23).

Proposition 11 (Energy) *Consider a solution $(n_{\sigma,\epsilon}, \varphi_{\sigma,\epsilon})$ of (2.24)–(2.25) defined by Theorem 10, then, the energy of the system $\mathcal{E}_{\sigma,\epsilon}$ satisfies*

$$\frac{d}{dt}\mathcal{E}_{\sigma,\epsilon}[n_{\sigma,\epsilon}(t)] = -\int_\Omega B_\epsilon(n_{\sigma,\epsilon})\left|\nabla(\varphi_{\sigma,\epsilon} + \psi'_{+,\epsilon}(n_{\sigma,\epsilon}))\right|^2 \leqslant 0. \tag{2.53}$$

As a consequence, we obtain a first a priori estimate

$$\mathcal{E}_{\sigma,\epsilon}[n_{\sigma,\epsilon}(T)] + \int_0^T \int_\Omega B_\epsilon(n_{\sigma,\epsilon})\big|\nabla(\varphi_{\sigma,\epsilon} + \psi'_{+,\epsilon}(n_{\sigma,\epsilon}))\big|^2 = \mathcal{E}_{\sigma,\epsilon}[n^0]. \tag{2.54}$$

Proof. To establish the energy of the regularized system, we begin with multiplying the first equation of (2.24) by $\varphi_{\sigma,\epsilon} + \psi'_{+,\epsilon}(n_{\sigma,\epsilon})$. Then, we integrate on the domain Ω and use the second boundary condition (2.25) to obtain

$$\int_\Omega [\varphi_{\sigma,\epsilon} + \psi'_{+,\epsilon}(n_{\sigma,\epsilon})]\partial_t n_{\sigma,\epsilon} = -\int_\Omega B_\epsilon(n_{\sigma,\epsilon})|\nabla(\varphi_{\sigma,\epsilon} + \psi'_{+,\epsilon}(n_{\sigma,\epsilon}))|^2.$$

Since $\psi'_{+,\epsilon}(n_{\sigma,\epsilon})\partial_t n_{\sigma,\epsilon} = \partial_t \psi_{+,\epsilon}(n_{\sigma,\epsilon})$, to retrieve the energy equality (2.53) we need to focus on the calculation of $\int_\Omega \varphi_{\sigma,\epsilon}\partial_t n_{\sigma,\epsilon}$. We write

$$\int_\Omega \varphi_{\sigma,\epsilon}\partial_t n_{\sigma,\epsilon} = \int_\Omega \varphi_{\sigma,\epsilon}\partial_t[n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}] + \frac{d}{dt}\int_\Omega \frac{\sigma}{2\gamma}|\varphi_{\sigma,\epsilon}|^2,$$

and using the second equation of (2.24), we rewrite the first term as

$$\begin{aligned}\int_\Omega \varphi_{\sigma,\epsilon}\partial_t[n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}] &= \int_\Omega [-\gamma\Delta(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}) + \psi'_-(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon})]\partial_t[n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}] \\ &= \frac{d}{dt}\int_\Omega \frac{\gamma}{2}\nabla(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon})|^2 + \psi_-(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}),\end{aligned}$$

where we have used the first boundary condition (2.25).

Altogether, we have recovered the expression (2.52) and the equality (2.53).

We can now turn to the entropy inequality. It is classical to define the mapping $\phi_\epsilon : [0,\infty) \mapsto [0,\infty)$

$$\phi''_\epsilon(n) = \frac{1}{B_\epsilon(n)}, \qquad \phi_\epsilon(0) = \phi'_\epsilon(0) = 0, \tag{2.55}$$

which is well defined because $B_\epsilon \in C(\mathbb{R},\mathbb{R}^+)$ from (2.20). For a nonnegative function $n(x)$, we define the entropy as

$$\Phi_\epsilon[n] = \int_\Omega \phi_\epsilon\big(n(x)\big)dx.$$

Proposition 12 (Entropy) *Consider a solution of (2.24)–(2.25) defined by Theorem 10, then the entropy of the system satisfies*

$$\begin{aligned}\frac{d\Phi_\epsilon[n_{\sigma,\epsilon}(t)]}{dt} = -\int_\Omega \gamma\left|\Delta\left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\right)\right|^2 + \frac{\sigma}{\gamma}|\nabla\varphi_{\sigma,\epsilon}|^2 + \psi''_-(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon})\left|\nabla(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon})\right|^2 \\ + \psi''_{+,\epsilon}(n_{\sigma,\epsilon})|\nabla n_{\sigma,\epsilon}|^2.\end{aligned} \tag{2.56}$$

Notice that the dissipation terms are all well defined by our definition of solution in Theorem 10. However, the equality (2.56) does not provide us with a direct a priori estimate because of the

negative term ψ''_-, therefore we have to combine it with the energy identity to write

$$\Phi_\epsilon[n_{\sigma,\epsilon}(T)] + \int_{\Omega_T} \left[\gamma \left|\Delta\left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\right)\right|^2 + \frac{\sigma}{\gamma}|\nabla\varphi_{\sigma,\epsilon}|^2 + \psi''_{+,\epsilon}(n_{\sigma,\epsilon})|\nabla n_{\sigma,\epsilon}|^2\right] \leqslant \Phi_\epsilon[n^0] + \frac{2T}{\gamma}\|\psi''_-\|_\infty\, \mathcal{E}_{\sigma,\epsilon}[n^0].$$

Proof. We compute, using the definition of ϕ''_ϵ,

$$\begin{aligned}
\int_\Omega \partial_t \phi_\epsilon(n_{\sigma,\epsilon}) &= \int_\Omega \partial_t n_{\sigma,\epsilon}\phi'_\epsilon(n_{\sigma,\epsilon}) \\
&= \int_\Omega \nabla\cdot\left[B_\epsilon(n_{\sigma,\epsilon})\nabla(\varphi_{\sigma,\epsilon} + \psi'_{+,\epsilon}(n_{\sigma,\epsilon}))\right]\phi'_\epsilon(n_{\sigma,\epsilon}) \\
&= -\int_\Omega B_\epsilon(n_{\sigma,\epsilon})\nabla(\varphi_{\sigma,\epsilon} + \psi'_{+,\epsilon}(n_{\sigma,\epsilon}))\phi''_\epsilon(n_{\sigma,\epsilon})\nabla n_{\sigma,\epsilon} \\
&= -\int_\Omega \nabla(\varphi_{\sigma,\epsilon} + \psi'_{+,\epsilon}(n_{\sigma,\epsilon}))\nabla n_{\sigma,\epsilon} \\
&= -\int_\Omega \nabla\varphi_{\sigma,\epsilon}\nabla(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}) + \psi''_{+,\epsilon}(n_{\sigma,\epsilon})|\nabla n_{\sigma,\epsilon}|^2 + \frac{\sigma}{\gamma}|\nabla\varphi_{\sigma,\epsilon}|^2.
\end{aligned} \tag{2.57}$$

To rewrite the term $\int_\Omega \nabla\varphi_{\sigma,\epsilon}\nabla(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon})$, we use the second equation of the regularized system (2.24)

$$\varphi_{\sigma,\epsilon} = -\gamma\Delta\left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\right) + \psi'_-(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}). \tag{2.58}$$

Using (2.58) and the boundary condition (2.25), we can rewrite the term under consideration as

$$\begin{aligned}
\int_\Omega \varphi_{\sigma,\epsilon}\Delta\left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\right) &= \int_\Omega -\gamma\left|\Delta\left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\right)\right|^2 + \psi'_-(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon})\Delta\left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\right) \\
&= -\int_\Omega \gamma\left|\Delta\left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\right)\right|^2 + \psi''_-(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon})\left|\nabla(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon})\right|^2.
\end{aligned}$$

Injecting this equality into (2.57), we obtain the identity (2.56).

2.2.4 Inequalities

From the energy and entropy properties, we can conclude the following a priori bounds, where we assume that the initial data has finite energy and entropy,

$$\frac{\sigma}{2\gamma}\int_\Omega |\varphi_{\sigma,\epsilon}(t)|^2 \leqslant \mathcal{E}_{\sigma,\epsilon}[n^0], \qquad \forall t \geqslant 0, \tag{2.59}$$

$$\frac{\sigma}{\gamma}\int_0^T\int_\Omega |\nabla\varphi_{\sigma,\epsilon}|^2 \leqslant \Phi_\epsilon[n^0] + \frac{2T}{\gamma}\|\psi''_-\|_\infty \mathcal{E}_{\sigma,\epsilon}[n^0], \qquad \forall T \geqslant 0, \tag{2.60}$$

$$\frac{\gamma}{2}\int_\Omega \left|\nabla\big(n_{\sigma,\epsilon}(t) - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}(t)\big)\right|^2 \leqslant \mathcal{E}_{\sigma,\epsilon}[n^0], \qquad \forall t \geqslant 0, \tag{2.61}$$

$$\int_0^T \int_\Omega \left| \Delta\left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\right)\right|^2 \leqslant \Phi_\epsilon[n^0] + \frac{2T}{\gamma}\|\psi_-''\|_\infty \mathcal{E}_{\sigma,\epsilon}[n^0], \qquad \forall T \geqslant 0, \tag{2.62}$$

$$\int_0^T \int_\Omega B_\epsilon(n_{\sigma,\epsilon})\left|\nabla(\varphi_{\sigma,\epsilon} + \psi'_{+,\epsilon}(n_{\sigma,\epsilon}))\right|^2 \leq \mathcal{E}_{\sigma,\epsilon}[n^0], \qquad \forall T \geqslant 0. \tag{2.63}$$

Proposition 13 (Compactness of time derivatives) *Consider a solution* $(n_{\sigma,\epsilon}, \varphi_{\sigma,\epsilon})$ *of* (2.24)–(2.25) *defined by Theorem 10, then, the following inequalities hold for* σ *small enough*

$$||\partial_t n_{\sigma,\epsilon}||_{L^2(0,T;(H^1(\Omega))')} \leq C, \tag{2.64}$$

$$||\partial_t \left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\right)||_{L^2(0,T;(H^1(\Omega))')} \leq C(\sigma). \tag{2.65}$$

Proof. For any test function $\chi \in L^2(0,T;H^1(\Omega))$ we obtained from (2.63)

$$\begin{aligned}\left|\int_{\Omega_T} \partial_t n_{\sigma,\epsilon}\chi\right| &= \left|\int_{\Omega_T} B_\epsilon(n_{\sigma,\epsilon})\nabla\left(\varphi_{\sigma,\epsilon} + \psi'_{+,\epsilon}(n_{\sigma,\epsilon})\right)\nabla\chi\right| \\ &\leq \left(\int_{\Omega_T} \left|B_\epsilon(n_{\sigma,\epsilon})\nabla\left(\varphi_{\sigma,\epsilon} + \psi'_{+,\epsilon}(n_{\sigma,\epsilon})\right)\right|^2\right)^{1/2} ||\nabla\chi||_{L^2(\Omega_T)}, \\ &\leq C||\nabla\chi||_{L^2(\Omega_T)}.\end{aligned}$$

This proves (2.64).

To prove (2.65), we compute the time derivative of equation for $\varphi_{\sigma,\epsilon}$ in the distribution sense

$$\sigma\int_{\Omega_T}\nabla U_{\sigma,\epsilon}\nabla\chi + \int_{\Omega_T} U_{\sigma,\epsilon}\chi = \int_{\Omega_T}\partial_t n_{\sigma,\epsilon}\chi - \frac{\sigma}{\gamma}\int_{\Omega_T} U_{\sigma,\epsilon}\psi_-''(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon})\chi,$$

where $U_{\sigma,\epsilon} = \partial_t\left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\right)$ and we have used the fact that $\left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\right)$, $n_{\sigma,\epsilon}$ and $\psi'_-(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon})$ are smooth. Then, we can choose $\chi = U_{\sigma,\epsilon}$, to obtain

$$\sigma\int_{\Omega_T}|\nabla U_{\sigma,\epsilon}|^2 + \int_{\Omega_T}|U_{\sigma,\epsilon}|^2 = \int_{\Omega_T}\partial_t n_{\sigma,\epsilon}U_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\int_{\Omega_T}|U_{\sigma,\epsilon}|^2\psi_-''(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}).$$

Using the fact that $\frac{\sigma}{\gamma}||\psi_-''||_\infty < 1$ from (2.7), the Cauchy-Schwarz inequality gives

$$\sigma||\nabla U_{\sigma,\epsilon}||^2_{L^2(\Omega_T)} + \alpha||U_{\sigma,\epsilon}||^2_{L^2(\Omega_T)} \leq ||\partial_t n_{\sigma,\epsilon}||_{L^2(\Omega_T)}||U_{\sigma,\epsilon}||_{L^2(\Omega_T)},$$

where $\alpha = 1 - \frac{\sigma}{\gamma}\|\psi_-''\|_\infty > 0$. Altogether, we obtain the bound (2.65) which is not uniform in σ.

2.3 Existence: convergence as $\epsilon \to 0$

The next step is to prove the existence of global weak solutions for the RDCH system (2.3) by letting ϵ vanish. This means that for all test functions $\chi \in L^2(0,T;H^1(\Omega)) \cap L^\infty(\Omega_T)$ with

$\nabla\chi \cdot \nu = 0$ on $\partial\Omega \times (0,T)$, it holds

$$\int_0^T < \chi, \partial_t n_\sigma > = \int_{\Omega_T} b(n_\sigma)\nabla\left(\varphi_\sigma + \psi'_+(n_\sigma)\right)\nabla\chi,$$
$$\sigma\int_{\Omega_T}\nabla\varphi_\sigma\nabla\chi + \int_{\Omega_T}\varphi_\sigma\chi = \gamma\int_{\Omega_T}\nabla n_\sigma\nabla\chi + \int_{\Omega_T}\psi'_-(n_\sigma - \frac{\sigma}{\gamma}\varphi_\sigma)\chi.$$

We establish the following

Theorem 14 (Existence for $\sigma > 0$, $\epsilon = 0$) *Assume an initial condition satisfying $0 \leqslant n^0 \leqslant 1$, with finite energy and entropy. Then, for σ small enough, there exists a global weak solution $(n_\sigma, \varphi_\sigma)$ of the RDCH equation (2.3), (2.4) such that*

$$n_\sigma \in L^2(0,T;H^1(\Omega)), \qquad \partial_t n_\sigma \in L^2\big(0,T;(H^1(\Omega))'\big). \tag{2.66}$$
$$\varphi_\sigma \in L^2(0,T;H^1(\Omega)), \tag{2.67}$$
$$n_\sigma - \frac{\sigma}{\gamma}\varphi_\sigma \in L^2(0,T;H^2(\Omega)), \qquad \partial_t\left(n_\sigma - \frac{\sigma}{\gamma}\varphi_\sigma\right) \in L^2\big(0,T;(H^1(\Omega))'\big). \tag{2.68}$$
$$0 \leq n_\sigma \leqslant 1, \qquad \text{a.e. in } \Omega_T, \tag{2.69}$$

and $n_\sigma < 1$ a.e. if b vanishes fast enough at 1 so that $\phi(1) = \infty$ (see (2.16)).

Proof. The proof relies on compactness results and the inequalities presented in section 2.2.4. From these inequalities, we can extract subsequences of $(n_{\sigma,\epsilon}, \varphi_{\sigma,\epsilon})$ such that the following convergences for $\epsilon \to 0$ hold for all $T > 0$.

Step 1. Weak limits. From (2.59) and (2.60), we immediately have

$$\varphi_{\sigma,\epsilon} \rightharpoonup \varphi_\sigma \text{ in } L^2\big((0,T);H^1(\Omega)\big). \tag{2.70}$$

Next, from (2.61), and the above convergence, we conclude

$$n_{\sigma,\epsilon} \rightharpoonup n_\sigma \text{ weakly in } L^2\big(0,T;H^1(\Omega)\big), \tag{2.71}$$

Finally from (2.64) and (2.65), we have

$$\partial_t n_{\sigma,\epsilon} \rightharpoonup \partial_t n_\sigma \text{ weakly in } L^2\big(0,T;(H^1(\Omega))'\big),$$
$$\partial_t\left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\right) \rightharpoonup \partial_t\left(n_\sigma - \frac{\sigma}{\gamma}\varphi_\sigma\right) \text{ weakly in } L^2\big(0,T;(H^1(\Omega))'\big).$$

Step 2. Strong convergence. Therefore, from the Lions-Aubin lemma and Proposition 13 we obtain the strong convergences

$$n_{\sigma,\epsilon} \to n_\sigma \in L^2(0,T;L^2(\Omega)). \tag{2.72}$$
$$n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon} \to n_\sigma - \frac{\sigma}{\gamma}\varphi_\sigma \in L^2(0,T;H^1(\Omega)). \tag{2.73}$$

Step 3. Bounds $0 \leqslant n_\sigma \leq 1$. To prove these bounds on n_σ, several authors have used the entropy relation. In the context of DCH equation with double-well potentials featuring singularities at $n = 1$ and $n = -1$, the solution lies a.e. in the interval $-1 < n < 1$. Elliott and Garcke [76] prove this result using the definition of the regularized entropy and by a contradiction argument. For single-well potential, Agosti *et al.* [8] used a reasoning on the measure of the set of solutions

outside the set $0 \leq n < 1$ and find contradictions with the boundedness of the entropy. This is the route we follow here. In the following, all functions are defined almost everywhere.

We begin by the upper bound . For $\alpha > 0$, we consider the set

$$V_\alpha^\epsilon = \{(t,x) \in \Omega_T | n_{\sigma,\epsilon}(t,x) \geq 1+\alpha\}.$$

For $A > 0$, there exists a small ϵ_0 such that the following estimate holds for every $\epsilon \leq \epsilon_0$

$$\phi_\epsilon''(n) = \frac{1}{b(1-\epsilon)} \geqslant 2A \qquad \forall n \geqslant 1,\ \forall \epsilon > 0.$$

Thus, integrating this quantity twice, we obtain

$$\phi_\epsilon(n) \geqslant A(n-1)^2 \qquad \forall n \geqslant 1.$$

Also, from (2.56), we know that the entropy is uniformly bounded in ϵ. Therefore, we obtain

$$|V_\alpha^\epsilon| A\alpha^2 \leqslant \int_{\Omega_T} \phi_\epsilon(n_{\sigma,\epsilon}(t,x)) \leqslant C(T), \qquad |V_\alpha^\epsilon| \leqslant \frac{C(T)}{A\alpha^2}.$$

In the limit $\epsilon \to 0$, using Fatou's lemma and the strong convergence of $n_{\sigma,\epsilon}$, we conclude that

$$\left|\{(t,x) \in \Omega_T | n_\sigma(t,x) \geq 1+\alpha\}\right| \leqslant \frac{C(T)}{A\alpha^2}, \qquad \forall A > 0.$$

In other words $n_\sigma(t,x) \leq 1+\alpha$ for all $\alpha > 0$, which means $n_\sigma(t,x) \leq 1$.

The same argument also gives $n_\sigma \geqslant 0$ and we do not repeat it.

The second statement, $n_\sigma < 1$ under the assumption $\phi(1) = +\infty$, is a consequence of the bound

$$\int_{\Omega_T} \phi(n_\sigma(t,x)) \leqslant C(T),$$

which holds true by strong convergence of $n_{\sigma,\epsilon}$ and because $\phi_\epsilon \nearrow \phi$ as $\epsilon \searrow 0$.

Step 4. Limiting equation. Finally, it remains to show that the limit of subsequences satisfies the RDCH equation in the weak form. Firstly, using the weak convergences (2.70)–(2.71), the strong convergence (2.73) and the properties of ψ_-' gathered from (2.7), we can pass to the limit in the standard way to obtain the second equation of the limit system.

To conclude the proof, we need to prove the following weak convergence, recalling that (2.63) provides a uniform L^2 bound over Ω_T, on $J_{\sigma,\epsilon}$

$$J_{\sigma,\epsilon} := -B_\epsilon(n_{\sigma,\epsilon})\nabla(\varphi_{\sigma,\epsilon} + \psi_{+,\epsilon}'(n_{\sigma,\epsilon})) \rightharpoonup -b(n_\sigma)\nabla(\varphi_\sigma + \psi_+'(n_\sigma)) \text{ weakly in } L^2(\Omega_T). \quad (2.74)$$

The convergence of $B_\epsilon(n_{\sigma,\epsilon})\nabla\varphi_{\sigma,\epsilon}$ follows from the weak convergence in $L^2(\Omega_T)$ of $\nabla\varphi_{\sigma,\epsilon}$ and the strong convergence $B_\epsilon(n_{\sigma,\epsilon}) \to b(n_\sigma)$ in all $L^p(\Omega_T)$, $1 \leqslant p < \infty$ which follows from (2.72) and the fact that $B_\epsilon(.) \to b(.)$ uniformly.

Because of the singularity $\psi_+'(1) = \infty$, we use the assumption (2.11) and that $B_\epsilon(\cdot)\psi_{+,\epsilon}''(\cdot) \to b(\cdot)\psi_+''(\cdot)$ uniformly and thus $B_\epsilon(n_{\sigma,\epsilon})\psi_{+,\epsilon}''(n_{\sigma,\epsilon}) \to b(n_{\sigma,\epsilon})\psi_+''(n_{\sigma,\epsilon})$ a.e. in Ω_T This achieve the proof.

It is easy to check that the energy and entropy relations (2.14), (2.17) hold, at least as inequalities. In the sequel we only use the a priori bounds coming from the limiting procedure.

2.4 Convergence as $\sigma \to 0$

We are now ready to study the limit of the relaxed solution n_σ towards a solution of the DCH equation, Our main result is as follows.

Theorem 15 (Limit $\sigma = 0$) *Let $(n_{\sigma,\epsilon}, \varphi_{\sigma,\epsilon})$ be a sequence of weak solutions of the RDHC system (2.24) with initial conditions n^0, $0 \leqslant n^0 < 1$, with finite energy and entropy. Then, as ϵ, $\sigma \to 0$, we can extract a subsequence of $(n_{\sigma,\epsilon}, \varphi_{\sigma,\epsilon})$ such that*

$$\varphi_{\sigma,\epsilon} \rightharpoonup -\gamma\Delta n + \psi'_-(n) \quad \text{weakly in } L^2(\Omega_T), \tag{2.75}$$

$$n_{\sigma,\gamma} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon} \to n \quad \text{strongly in } L^2(0,T;H^1(\Omega)), \tag{2.76}$$

$$n_{\sigma,\epsilon},\ \nabla n_{\sigma,\epsilon} \to n,\ \nabla n \quad \text{strongly in } L^2(\Omega_T), \text{ and } 0 \leqslant n \leqslant 1, \tag{2.77}$$

and $n_\sigma < 1$ a.e. if b vanishes fast enough at 1 so that $\phi(1) = \infty$.

$$\partial_t n_{\sigma,\epsilon} \rightharpoonup \partial_t n \text{ weakly in } L^2\big(0,T;(H^1(\Omega))'\big). \tag{2.78}$$

This limit n satisfies the DCH system (2.1) in the weak sense.

We recall the definition of weak solutions; for all $\chi \in L^2(0,T;H^2(\Omega)) \cap L^\infty(\Omega_T)$ with $\nabla\chi\cdot\nu = 0$ on $\partial\Omega \times (0,T)$,

$$\begin{cases} \int_0^T < \chi, \partial_t n > &= \int_{\Omega_T} J \cdot \nabla\chi, \\ \int_{\Omega_T} J \cdot \nabla\chi &= -\int_{\Omega_T} \gamma\Delta n \left[b'(n)\nabla n \cdot \nabla\chi + b(n)\Delta\chi\right] + (b\psi'')(n)\nabla n \cdot \nabla\chi. \end{cases} \tag{2.79}$$

Proof. We gathered, from the energy and entropy estimates of section 2.2.3, the a priori bounds of the section 2.2.4.

Step 1. Weak limits. From the above mentioned inequalities, we can extract subsequences of $(n_{\sigma,\epsilon}, \varphi_{\sigma,\epsilon})$ such that the following convergences hold for all $T > 0$. From (2.59) and (2.60), we immediately have

$$\sigma\varphi_{\sigma,\epsilon} \to 0 \text{ in } L^2\big((0,T);H^1(\Omega)\big). \tag{2.80}$$

Next, from (2.61), and the above convergence, we conclude

$$n_{\sigma,\epsilon} \rightharpoonup n \text{ weakly in } L^2\big(0,T;H^1(\Omega)\big),$$

and (2.62) gives directly

$$\Delta(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}) \rightharpoonup \Delta n \text{ weakly in } L^2(\Omega_T). \tag{2.81}$$

This latter convergence is obtained in the distribution sense using integration per parts, for all test function $\chi \in \mathcal{D}(\Omega_T)$

$$\int_{\Omega_T} \Delta\big(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\big)\chi = -\int_{\Omega_T} \nabla\big(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\big)\nabla\chi.$$

Then using (2.80), we obtain (2.81). The system of equations can also be used to complement these results. We find

$$\varphi_{\sigma,\epsilon} \rightharpoonup \varphi \text{ weakly in } L^2(\Omega_T),$$

using the second equation of the system (2.24) and triangular inequality,

$$\|\varphi_{\sigma,\epsilon}\|_{L^2(\Omega_T)} \leq \gamma\|\Delta(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon})\|_{L^2(\Omega_T)} + \|\psi'_-(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon})\|_{L^2(\Omega_T)}.$$

Finally from (2.63) and the equation on $n_{\sigma,\epsilon}$ itself, we conclude (2.78).

Step 2. Strong convergence. We continue with proving the strong convergences in (2.77). From the inequality (2.62), we know that $\Delta\left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\right)$ is uniformly bounded in $L^2(\Omega_T)$. We also have the boundary conditions, $\nabla\left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi\right)\cdot\nu = 0$ and the conservation of both quantities. Therefore elliptic regularity theory gives us

$$\|n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\|_{L^2(0,T;H^2(\Omega))} \leq C.$$

Therefore strong compactness in space holds for the quantities $n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi$ and $\nabla[n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi]$. Furthermore, from the limit (2.80), it means that both $n_{\sigma,\epsilon}$ and $\nabla n_{\sigma,\epsilon}$ are compact in space. Compactness in time is also obtained for the quantity $n_{\sigma,\epsilon}$ from (2.64). Again from Lions-Aubin lemma, we have the strong convergence (2.77). The conclusions (2.76) and (2.75) follows from this results.

The bounds $0 \leqslant n < 1$ can be obtained as in the case $\epsilon \to 0$, see Theorem 14 and we do not repeat the argument.

Step 3. Limiting equation. Next, we need to verify that the limit of the subsequence $n_{\sigma,\epsilon}$ satisfies the DCH equation. The argument is different from the case $\epsilon \to 0$ because we do not control $\nabla\varphi_{\sigma,\epsilon}$ in the case at hand. From the L^2 bound in (2.63), we need to identify the weak limit

$$J_{\sigma,\epsilon} := -B_\epsilon(n_{\sigma,\epsilon})\nabla(\varphi_{\sigma,\epsilon} + \psi'_{+,\epsilon}(n_{\sigma,\epsilon})) \rightharpoonup -b(n)\nabla(\varphi + \psi'_+(n)) \quad \text{weakly in } L^2(\Omega_T). \tag{2.82}$$

For a test function $\eta \in L^2(0,T;H^1(\Omega,\mathbb{R}^d)) \cap L^\infty(\Omega_T,\mathbb{R}^d)$ and $\eta\cdot\mu = 0$ on $\partial\Omega\times(0,T)$, we integrate the left-hand side to obtain

$$\begin{aligned}\int_{\Omega_T} J_{\sigma,\epsilon}\cdot\eta = -\int_{\Omega_T}\Big[&\gamma\Delta\left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\right)\nabla\cdot(B_\epsilon(n_{\sigma,\epsilon})\eta)\\ &+ B_\epsilon(n_{\sigma,\epsilon})\nabla\left(\psi'_{+,\epsilon}(n_{\sigma,\epsilon}) + \psi'_-(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon})\right)\cdot\eta\Big].\end{aligned}$$

We have mainly two types of terms on the right-hand side $\int_{\Omega_T}\gamma\Delta\left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\right)\nabla\cdot(B_\epsilon(n_{\sigma,\epsilon})\eta)$ and $\int_{\Omega_T} B_\epsilon(n_{\sigma,\epsilon})\nabla\left(\psi'_{+,\epsilon}(n_{\sigma,\epsilon}) + \psi'_-(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon})\right)\cdot\eta$. Let us focus on the first term

$$\begin{aligned}\int_{\Omega_T}\gamma\Delta\left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\right)\nabla\cdot(B_\epsilon(n_{\sigma,\epsilon})\eta) = &\int_{\Omega_T}\gamma\Delta\left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\right)B_\epsilon(n_{\sigma,\epsilon})\nabla\cdot\eta\\ &+\int_{\Omega_T}\gamma\Delta\left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\right)B'_\epsilon(n_{\sigma,\epsilon})\nabla n_{\sigma,\epsilon}\cdot\eta.\end{aligned}$$

From the strong convergence (2.77) and the weak one (2.81) with the fact that $B_\epsilon(\cdot) \to b(\cdot)$ uniformly, we obtain the convergence of the first term of the right-hand side

$$\int_{\Omega_T}\gamma\Delta\left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\right)B_\epsilon(n_{\sigma,\epsilon})\nabla\cdot\eta \to \int_{\Omega_T}\gamma\Delta n\, b(n)\nabla\cdot\eta,$$

as $\sigma, \epsilon \to 0$ and thus we have passed to the limit in the first term of the right hand side. For the second term, we use that the derivative $B'_\epsilon(\cdot) \to b'(\cdot)$ uniformly. We also use the strong convergence of $\nabla n_{\sigma,\epsilon}$ from (2.77). From the results above and a generalized version of the Lebesgue dominated convergence theorem we obtain

$$\int_{\Omega_T} \gamma\Delta\left(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon}\right) B'_\epsilon(n_{\sigma,\epsilon})\nabla n_{\sigma,\epsilon}\cdot\eta \to \int_{\Omega_T} \gamma\Delta n b'(n)\nabla n\cdot\eta,$$

as $\sigma, \epsilon \to 0$.

Let us now pass to the limit in $\int_{\Omega_T} B_\epsilon(n_{\sigma,\epsilon})\nabla\left(\psi'_{+,\epsilon}(n_{\sigma,\epsilon}) + \psi'_-(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon})\right)\cdot\eta$. As in the case of the convergence $\epsilon \to 0$, we have that

$$\int_{\Omega_T} B_\epsilon(n_{\sigma,\epsilon})\nabla\left(\psi'_{+,\epsilon}(n_{\sigma,\epsilon})\right)\cdot\eta,$$

using the fact that $B_\epsilon(\cdot)\psi''_{+,\epsilon}(\cdot) \to b(\cdot)\psi''_+(\cdot)$ uniformly and the strong convergence (2.77). Since $B_\epsilon(\cdot) \to b(\cdot)$, we have $\left(B_\epsilon\psi''_-\right)(\cdot) \to \left(b\psi''_-\right)(\cdot)$.
Therefore, we pass to pass to the limit in $\int_{\Omega_T} B_\epsilon(n_{\sigma,\epsilon})\nabla\left(\psi'_-(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon})\right)\cdot\eta$ using the convergence (2.76). Altogether, we obtain the following convergence

$$\int_{\Omega_T} B_\epsilon(n_{\sigma,\epsilon})\nabla\left(\psi'_{+,\epsilon}(n_{\sigma,\epsilon}) + \psi'_-(n_{\sigma,\epsilon} - \frac{\sigma}{\gamma}\varphi_{\sigma,\epsilon})\right)\cdot\eta \to \int_{\Omega_T} b(n)\nabla\left(\psi'_+(n) + \psi'_-(n)\right)\cdot\eta$$

This finishes the proof of (2.82), i.e. that the limit solution n satisfies the weak formulation of the DCH equation (2.1), and also the proof of Theorem 15.

2.5 Long-time behavior

To complete our study of the RDCH model, we give some insights concerning the long-time behavior and convergence to steady states, $(n_\infty, \varphi_\infty)$ determined by the steady problem

$$\begin{cases} \nabla\cdot\left(b(n_\infty)\nabla\left(\varphi_\infty + \psi'_+(n_\infty)\right)\right) = 0 \quad \text{in } \Omega, \\ -\sigma\Delta\varphi_\infty + \varphi_\infty = -\gamma\Delta n_\infty + \psi'_-(n_\infty - \frac{\sigma}{\gamma}\varphi_\infty) \text{ in } \Omega, \\ \frac{\partial\left(n_\infty - \frac{\sigma}{\gamma}\varphi_\infty\right)}{\partial\nu} = b(n_\infty)\frac{\partial\left(\varphi_\infty + \psi'_+(n_\infty)\right)}{\partial\nu} = 0 \quad \text{on } \partial\Omega. \end{cases} \tag{2.83}$$

The analysis of the steady-states is not performed in this paper, however, numerical simulations can help us to have an idea of their shape for different initial situations.

The steady-states of the RDCH model present a configuration which minimizes the energy of the system. The solution obtained at the end of the simulation depends mainly on three parameters: the initial mass M, the width of the diffuse interface $\sqrt{\gamma}$ and the relaxation parameter σ.

In fact, if the initial mass is large enough, saturated aggregates are formed and we can describe two regions in the domain: the aggregates and the absence of cells. Between these two regions, the transition is smooth and the length of this interface is $\sqrt{\gamma}$. If the initial mass is small, aggregates are still formed but they are thicker and their maximum concentration does not reach 1 or the critical value $n^\star$ as in the definition of the potential (2.8).

The formation of aggregates happens only if γ is small enough. If γ, the initial mass M or

the relaxation parameter σ is too large, the solution converges to the constant one

$$n_\infty = \frac{1}{|\Omega|}\int_\Omega n^0 dx, \quad \text{a.e. in } \Omega.$$

A surprizing fact about these observations is that the long-time behavior of the solutions of the RDCH system seems to follow the analytical description of the steady-states made by Songmu [183].

To state our convergence result of the weak solutions of the RDCH model to steady-states, we consider a global weak solution (n, φ) of the RDCH system with $\sigma > 0$, according to Theorem 14. The initial condition satisfies $0 \le n^0 < 1$ and has finite energy and entropy. so that we can use the a priori estimates from the transport structure, the energy and entropy dissipations (2.53) and (2.56) (or (2.14)–(2.17)), in particular

$$0 \le n < 1 \text{ a.e. } (0,\infty)\times\Omega. \tag{2.84}$$

Based on the controls provided by these relations, and using a standard method, we are going to study the large time behavior as the limit for large k of the sequence of functions

$$n_k(t,x) = n(t+k,x), \qquad \text{and} \qquad \varphi_k(t,x) = \varphi(t+k,x).$$

Proposition 16 (Long term convergence along subsequences) *Let (n,φ) be a weak solution of (2.3), (2.4) and initial condition n^0 with $0 \le n^0 < 1$, finite energy and entropy. Then, we can extract a subsequence, still denoted by index k, of (n_k, φ_k) such that*

$$\lim_{k\to\infty} n_k(x,t) = n_\infty(x), \quad \lim_{k\to\infty} \varphi_k(x,t) = \varphi_\infty(x) \qquad \text{strongly in } L^2\big((-T,T)\times\Omega\big), \quad \forall T>0, \tag{2.85}$$

where $(n_\infty, \varphi_\infty)$ are solutions of (2.83) satisfying

$$b(n_\infty)\nabla\left(\varphi_\infty + \psi'_+(n_\infty)\right) = 0. \tag{2.86}$$

Proof. The proof uses the energy and entropy inequalities to obtain both uniform (in k) a priori bounds and zero entropy dissipation in the limit, which imply the result. We write these arguments in several steps.

1st step. A priori bounds from energy. Energy decay implies that $\mathcal{E}[n_k(t)]$ remains bounded in k for $t > -k$. As a consequence, the sequence (n_k, φ_k) satisfies

$$\frac{\sigma}{2\gamma}\int_\Omega |\varphi_k(t)|^2 \leqslant \mathcal{E}[n^0], \qquad \forall t \geqslant 0, \tag{2.87}$$

$$\frac{\gamma}{2}\int_\Omega \left|\nabla\big(n_k(t) - \frac{\sigma}{\gamma}\varphi_k(t)\big)\right|^2 \leqslant \mathcal{E}[n^0], \qquad \forall t \geqslant 0, \tag{2.88}$$

$$\int_{-T}^{T}\int_\Omega b(n_k)\big|\nabla(\varphi_k + \psi'_+(n_k))\big|^2 := L_k(T), \qquad L_k(T)\to 0 \ \text{ as } \ k\to\infty, \tag{2.89}$$

and this last line is because

$$\int_0^\infty\int_\Omega b(n)\big|\nabla(\varphi+\psi'_+(n))\big|^2 \le \mathcal{E}[n^0], \quad L_k(T) \leqslant \int_{k-T}^\infty\int_\Omega b(n)\big|\nabla(\varphi+\psi'_+(n))\big|^2] \underset{k\to\infty}{\longrightarrow} 0.$$

2nd step. A priori bounds from entropy. Because the right hand side in the entropy balance has a positive term (since $\psi_-''(n) \le 0, \forall n \in [0,1]$), it cannot be used as easily as the energy. However, we can integrate (2.17) from $k-T$ to $k+T$, and, using the control of the negative term including ψ_- as after (2.17), we obtain the inequality

$$\begin{aligned}
&\int_{-T}^{T} \left[\left| \Delta \left(n_k - \frac{\sigma}{\gamma}\varphi_k \right) \right|^2 + \frac{\sigma}{\gamma} |\nabla \varphi_k|^2 + \psi_+''(n_k) |\nabla n_k|^2 \right] \\
&\quad \le \Phi[n(k-T)] - \Phi[n(k+T)] + \|\psi_-''\|_\infty \left\| \nabla \left(n_k - \frac{\sigma}{\gamma}\varphi_k \right) \right\|^2_{L^2((-T,T)\times\Omega)} \\
&\quad \le \Phi[n(k-T)] - \Phi[n(k+T)] + \|\psi_-''\|_\infty \frac{4T}{\gamma} \mathcal{E}[n(k-T)] \\
&\quad \le \Phi[n(k-T)] - \Phi[n(k+T)] + \|\psi_-''\|_\infty \frac{4T}{\gamma} \mathcal{E}[n^0].
\end{aligned}$$

3rd step. Extracting subsequences. From these inequalities, we can extract subsequences of (n_k, φ_k) such that for $k \to \infty$, the following convergences hold toward some functions $n_\infty(x,t)$ and $\varphi_\infty(x,t)$.

We can conclude from inequalities (2.87) and the entropy control that, as $k \to \infty$,

$$\varphi_k \rightharpoonup \varphi_\infty \text{ weakly in } L^2\big(-T, T; H^1(\Omega)\big). \tag{2.90}$$

From the gradient bound (2.88), the L^2 bound in (2.87) and $0 \leqslant n_k < 1$, we obtain

$$n_k - \frac{\sigma}{\gamma}\varphi_k \rightharpoonup n_\infty - \frac{\sigma}{\gamma}\varphi_\infty \text{ weakly in } L^2\big(0, T; H^1(\Omega)\big), \tag{2.91}$$

and thus

$$n_k \rightharpoonup n_\infty \text{ weakly in } L^2\big(0, T; H^1(\Omega)\big). \tag{2.92}$$

Finally, we obtain from (2.89) and the Cauchy-Schwarz inequality,

$$\partial_t n_k \rightharpoonup \partial_t n_\infty = 0 \text{ weakly in } L^2\big(0, T; (H^1(\Omega))'\big). \tag{2.93}$$

Indeed, for any test function $\phi \in C_0^\infty((-T,T)\times\Omega)$, it holds

$$\int_{-T}^{T}\int_\Omega \partial_t n_k \phi dx dt = -\int_{-T}^{T}\int_\Omega b(n_k) \nabla \left(\varphi_k + \psi_+'(n_k) \right) \cdot \nabla \phi,$$

$$\left| \int_{-T}^{T}\int_\Omega \partial_t n_k \phi dx dt \right|^2 \leqslant 2T|\Omega| \|b\|_\infty \|\nabla \phi\|_\infty^2 \int_{-T}^{T}\int_\Omega b(n_k) \left| \nabla \left(\varphi_k + \psi_+'(n_k) \right) \right|^2 \to 0$$

as $k \to \infty$. This also shows that n_∞ only depends on x.

4th step. Strong limits. The strong compactness of n_k and φ_k follows from (2.88) and the entropy control. Then, time compactness of n_k, stated in (2.85) follows from the Lions-Aubin lemma, thanks to (2.93). The strong convergence of φ_k is a consequence of the elliptic equation for φ_k and of (2.65) which gives compactness in time of the quantity $n_k - \frac{\sigma}{\gamma}\varphi_k$. And we also have, from the strong convergence of n_k and (2.91), thanks to the above argument,

$$b(n_k) \nabla \left(\varphi_k + \psi_+'(n_k) \right) \to b(n_\infty) \nabla \left(\varphi_\infty + \psi_+'(n_\infty) \right) = 0, \tag{2.94}$$

which establishes the zero-flux equality (2.86).

2.6 Conclusion

The proposed relaxation system of the degenerate Cahn-Hilliard equation with single-well potential reduces the model to two parabolic/elliptic equations which can be solved by standard numerical solvers. The relaxation uses a regularization in space of the new unknown used to transform the original fourth-order equation into two second-order equations. This new system is a non-local relaxation of the original equation which is similar in a sense to the Cahn-Hilliard equation with a spatial interaction kernel derived in [99, 100]. We proved that in the limit of vanishing relaxation, we retrieve the original weak solutions of the DCH equation using compactness methods and estimates borrowed from energy and entropy functionals. The long-time behavior of the solutions of the RDCH system can also be studied along the same lines. We showed that a global solution of the system converges to a steady-state as time goes to infinity, with zero flux.

The stationary states exhibit some interesting properties due to the degeneracy of the mobility. More precisely, they are split into two distinct zones: whether the mobility is zero, which is possible only in the pure phases, or the flux is null.

The RDCH system aims at the design of a numerical method to simulate the DCH equation using only second order elliptic problems. Such a numerical scheme may depend on details of the relaxed model. For example, the solution represents a density and its numerical positivity is a desired property. Also, the discrete stability is useful and a change of unknown in the RDCH system might be better adapted, using $U = \varphi - \frac{\gamma}{\sigma} n$,

$$\begin{aligned} \partial_t n &= \nabla \cdot \left(b(n) \nabla \left(U + \frac{\gamma}{\sigma} n + \psi_+'(n) \right) \right), \\ -\sigma \Delta U + U &= -\frac{\gamma}{\sigma} n + \psi_-'(-\frac{\sigma}{\gamma} U). \end{aligned}$$

Even though this model also consists of a parabolic transport equation coupled with an elliptic equation, the regularity is enhanced. On the one hand, in the first equation, the term $\frac{\gamma}{\sigma} n$ increases the diffusion for n. On the other hand, the second equation regularizes for the new variable U because it depends on n rather than Δn. In a forthcoming work, we will propose a numerical scheme based on the RDCH system, that preserves the physical properties of the solutions.

Chapter 3

Structure-preserving numerical method for the relaxed-degenerate Cahn-Hilliard model

Abstract

We propose and analyze two finite element approximations of the relaxed Cahn-Hilliard equation [165] with singular single-well potential of Lennard-Jones type and degenerate mobility that are energy stable and nonnegativity preserving. The Cahn-Hilliard model has recently been applied to model evolution and growth for living tissues: although the choices of degenerate mobility and singular potential are biologically relevant, they induce difficulties regarding the design of a numerical scheme. We propose finite element schemes and we show that they preserve the physical bounds of the solutions thanks to two different suitable approximations of the mobility. Indeed, in the first scheme, the mobility is approximated by a piecewise constant matrix on each element of the mesh, and for the second one, we propose an adaptation of the upwind method to improve the efficiency of the simulations. Moreover, we analyze well-posedness, energy stability properties, and convergence of solutions for the different numerical schemes. Finally, we validate our numerical method by presenting numerical simulations in one and two dimensions.

This chapter is taken from Federica Bubba, A. P., *A nonnegativity preserving scheme for the relaxed Cahn-Hilliard equation with single-well potential and degenerate mobility*, (2020), submitted for publication.

3.1 Introduction

Being of fourth order, the Cahn-Hilliard equation does not fit usual softwares for finite elements. To circumvent this difficulty a relaxed version has been proposed in [165] and the presentation of two finite element numerical schemes that preserve the physical properties of the solutions is the purpose of the present work. The relaxed version of the Cahn-Hilliard equation reads

$$\begin{cases} \dfrac{\partial n}{\partial t} = \nabla \cdot \left(b(n) \nabla \left(\varphi + \psi'_+(n) \right) \right), \\ \\ -\sigma \, \Delta \varphi + \varphi = -\gamma \, \Delta n + \psi'_- \left(n - \dfrac{\sigma}{\gamma} \varphi \right), \end{cases} \qquad t > 0, \, x \in \Omega, \tag{3.1}$$

and is set in a regular bounded domain $\Omega \subset \mathbb{R}^d$ with $d = 1, 2, 3$. It describes the evolution in time of the (relative) volume fraction $n \equiv n(t,x)$ of one of the two components in a binary mixture. The system is equipped with nonnegative initial data

$$n(0,x) = n^0(x) \in H^1(\Omega), \qquad 0 \leq n^0(x) < 1, \quad x \in \Omega,$$

and with zero-flux boundary conditions on the boundary $\partial\Omega$ of Ω

$$\frac{\partial(n - \frac{\sigma}{\gamma}\varphi)}{\partial\nu} = b(n)\,\frac{\partial\left(\varphi + \psi'_+(n)\right)}{\partial\nu} = 0, \quad t > 0,\, x \in \partial\Omega,$$

where ν is the unit normal vector pointing outward $\partial\Omega$.

System (3.1) was proposed in [165] as an approximation, in the asymptotic regime whereby the *relaxation parameter* σ vanishes (*i.e.*, $\sigma \to 0$), of the fourth order Cahn-Hilliard equation [47, 48]. The Cahn-Hilliard (CH) equation describes spinodal decomposition phenomena occurring in binary alloys after quenching: an initially uniform mixed distribution of the alloy undergoes phase separation and a two-phase inhomogeneous structure arises. In its original form, the Cahn-Hilliard equation is written in the form of an evolution equation for n:

$$\partial_t n = \nabla \cdot \left(b(n)\nabla\left(\psi'(n) - \gamma\Delta n\right)\right), \quad t > 0,\, x \in \Omega, \tag{3.2}$$

with $n \in [-1, 1]$, where the states $n \equiv -1$ and $n \equiv 1$ denote the two pure phases arising after the mixture has undergone the phase separation process. Writing the flux as $\mathbf{J} = -b(n)\nabla\left(\frac{\delta\mathcal{E}[n]}{\delta n}\right)$, Equation (3.2) can be interpreted as the conservative gradient flow of the free energy functional

$$\mathcal{E}[n](t) := \int_\Omega \left(\frac{\gamma}{2}\left|\nabla n\right|^2 + \psi(n)\right) \mathrm{d}x.$$

The *homogeneous free energy* ψ describes repulsive and attractive interactions between the two components of the mixture while the regularizing term $\frac{\gamma}{2}\left|\nabla n\right|^2$ accounts for partial mixing between the pure phases, leading to a *diffuse interface* separating the states $n \equiv -1$ and $n \equiv 1$, of thickness proportional to $\sqrt{\gamma}$. The parameter $\gamma > 0$ is related to the surface tension at the interface (see, *e.g.*, [146]) and the function b is called *mobility*.
In most of the literature, ψ is a double-well logarithmic potential, often approximated by a smooth polynomial function, with minimums located at the two attraction points that represent pure phases $n = \pm 1$ (see, *e.g.*, [60, 77, 75]). The mobility can be either constant [77, 75] or degenerate at the pure phases [28, 76]. We refer to the introductory chapters [74, 154] and to the recent review [146] for an overview of the derivation of the Cahn-Hilliard equation, its analytical properties and its variants.

Recently, the Cahn-Hilliard equation has been considered as a phenomenological model for the description of cancer growth; see, for instance, [6, 55, 203]. In this context, n represents the volume fraction of the tumor in a two-phase mixture containing cancerous cells and a liquid phase, such as water and other nutrients. In biological contexts, a double-well potential appears to be nonphysical. In fact, as suggested by Byrne and Preziosi in [46], a single-well potential of Lennard-Jones type allows for a more suitable description of attractive and repulsive forces acting in the mixture. Following this intuition and building upon previous works [8, 55], in this paper we consider a single-well homogeneous free energy $\psi : [0, 1) \to \mathbb{R}$, defined as

$$\psi(n) = -(1 - n^\star)\log(1 - n) - \frac{n^3}{3} - (1 - n^\star)\frac{n^2}{2} - (1 - n^\star)n + k, \quad n^\star > 0, \tag{3.3}$$

where $k \in \mathbb{R}$ is an arbitrary constant. In the above form, ψ models cell-cell attraction at small densities ($\psi'(\cdot) < 0$ for $0 < n \leqslant n^\star$ and $\psi'(0) = 0$) and repulsion in overcrowded zones ($\psi'(\cdot) > 0$ for $n \geqslant n^\star$); *cf.* Figure 3.1. The quantity $n^\star > 0$ represents the value of the cellular density at which repulsive and attractive forces are at equilibrium. With a potential of the form (3.3), the pure phases are represented by the states $n = 0$ and $n = 1$, where $n = 1$ is a singularity for ψ is such a way to avoid overcrowding. Moreover, we consider a degenerate mobility b, that has to vanish at $n = 0$ and $n = 1$. For instance, as in [8], we choose

$$b(n) := n\,(1-n)^2. \tag{3.4}$$

The Cahn-Hilliard equation (3.2) with the logarithmic single-well potential defined in (3.3) and a mobility given by (3.4) has been studied by Agosti *et al.* in [8], where the authors prove well-posedness of the equation for $d \leq 3$.

Summary of previous results and specific difficulties. Numerous numerical methods have been developed to solve the Cahn-Hilliard equation (3.2) with smooth and/or logarithmic double-well potential as well as with constant or degenerate mobility. Generally, a numerical scheme for the Cahn-Hilliard equation is evaluated by several aspects: *i)* its capacity to keep the energy dissipation (energy stability) and the physical bounds of the solutions; *ii)* if it is convergent, and if error bounds can be established; *iii)* its efficiency; *iv)* its implementation simplicity. To meet the first point concerning the energy stability, several implicit schemes have been proposed. The main drawback of these methods is the necessity to use an iterative method to solve the resulting nonlinear system. To circumvent this issue, unconditionally energy-stable schemes have been proposed based on the splitting of the potential in a convex and a non-convex part. This idea comes from Eyre [82] and leads to unconditionally energy-stable explicit-implicit (i.e. semi-implicit) approximations of the model. For references on all the previous numerical methods discussed above, we refer the reader to the review paper [193].

For finite element approximations, most of these results are based on the second-order splitting

$$\begin{cases} \partial_t n = \nabla \cdot (b(n) \nabla w)\,, \\ w = -\gamma \Delta n + \psi'(n), \end{cases} \tag{3.5}$$

where, w is called chemical potential; see, *e.g.*, [8, 28, 75].
In [77], Elliot and Songmu propose a finite element Galerkin approximation for the resolution of (3.2) with a smooth double-well potential and constant mobility. The more challenging case of a degenerate mobility and singular potentials has been considered by Barrett *et al.* in [28], where the authors propose a finite element approximation which employs the second-order splitting (3.5). In particular, the authors provide well-posedness of the finite element approximation as well as a convergence result in the one-dimensional case. Numerical methods to solve the Cahn-Hilliard equation without the splitting technique (3.5) have also been suggested. For instance, in [42] Brenner *et al.* propose a C^0 interior penalty method, a class of discontinuous Galerkin-type approximations.

Even though a single-well potential seems more relevant for biological applications of the Cahn-Hilliard equation, very few works focus on this case. In the already mentioned [8], Agosti and collaborators propose a finite element method to solve Equation (3.2) with the homogeneous energy given by (3.3) and a degenerate mobility of the form (3.4). As the authors remark, the main issues arising when considering a single-well logarithmic potential is that the positivity of the solution is not ensured at the discrete level, since the mobility degeneracy set $\{0; 1\}$ does not coincide with the singularity set of the potential, *i.e.*, $n = 1$. Therefore, the absence of

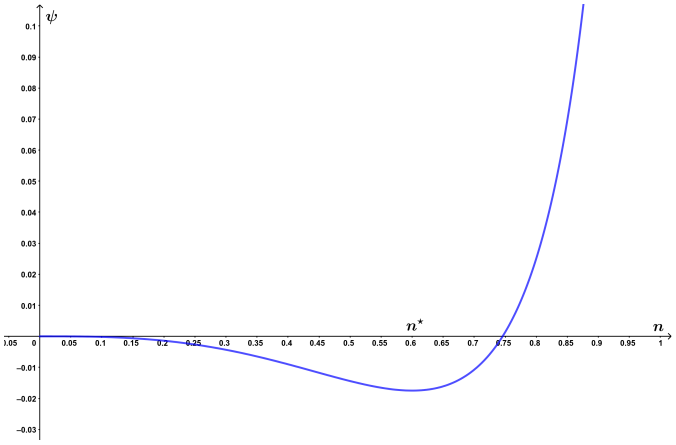


Figure 3.1 – Single-well potential of Lennard-Jones type as in (3.6) with $n^\star = 0.6$.

cells represents an unstable equilibrium of the potential. In [8], the authors design a finite element scheme which preserves positivity by the means of a discrete variational inequality, as also suggested in [28]. More recently, in [7], Agosti has presented a discontinuous Galerkin finite element discretization of the equation where, again, the positivity of the discrete solution is ensured thanks to a discrete variational inequality.

Contents of the paper. The aim of this paper is to present and analyze two finite element approximations of the relaxed Cahn-Hilliard equation (3.1) with single-well potential (3.3) and degenerate mobility (3.4) in dimensions $d = 1, 2, 3$. More in details, we prove: (i) well-posedness of the numerical approximation; (ii) nonnegativity of discrete solutions ensured by a suitable approximation of the mobility function b; (iii) discrete energy and entropy bounds; (iv) convergence of discrete solutions in dimension $d = 1$.
In System (3.1), ψ_+ and ψ_- are, respectively, the convex and concave part of ψ, defined as

$$\psi_+(n) := -(1-n^\star)\log(1-n) - \frac{n^3}{3}, \quad \text{and} \quad \psi_-(n) := -(1-n^\star)\frac{n^2}{2} - (1-n^\star)n, \tag{3.6}$$

where ψ_+ is convex whenever $n^\star \leq 1 - \left(\frac{2}{3}\right)^3$.

The main novelty of our work is to propose an alternative to the second-order splitting (3.5) by replacing the chemical potential w by its relaxed approximation φ, solution to a second order elliptic equation with diffusivity $0 < \sigma \ll \gamma$. The relaxed system is based on the analysis performed in [165], where the authors prove well-posedness of the system as well as the convergence, as $\sigma \to 0$, of weak solutions of (3.1) to the ones of the original Cahn-Hilliard equation (3.2). For the analysis that follows, it is worth noticing that System (3.1) admits the energy functional

$$\mathcal{E}_\sigma[n](t) := \int_\Omega \left\{ \frac{\gamma}{2}\left|\nabla\left(n - \frac{\sigma}{\gamma}\varphi\right)\right|^2 + \frac{\sigma}{2\gamma}|\varphi|^2 + \psi_+(n) + \psi_-\left(n - \frac{\sigma}{\gamma}\varphi\right)\right\} \mathrm{d}x, \tag{3.7}$$

that, as proved in [165], is decreasing in time, *i.e.*,

$$\frac{\mathrm{d}\mathcal{E}_\sigma[n]}{\mathrm{d}t} = -\int_\Omega b(n)\left|\nabla\left(\varphi + \psi_+'(n)\right)\right|^2 \mathrm{d}x \leqslant 0, \quad t > 0.$$

We also notice that the convex/concave splitting of ψ is different from the one employed, *e.g.*, in [8] and is motivated by the need to retrieve energy dissipation as well as by the fact that we can take advantage of the linearity of ψ'_- to achieve regularity results on φ. Furthermore, we observe that the relaxed Cahn-Hilliard system bears some similarities with the Keller-Segel model with additional cross diffusion, proposed and analyzed in [32, 52].

In this work we aim to describe two finite element schemes that preserve the physical bounds of the solutions of the relaxed-degenerate Cahn-Hilliard model (RDCH in short).

This paper is organized as follows. We start in Section 3.2 by giving details about the finite element framework we are using. Then in Section 3.3, we recall the regularized version of the relaxed-degenerate Cahn-Hilliard model (3.1). In Section 3.4, we introduce a non-linear semi-implicit finite element approximation of the regularized-relaxed model. The definition of the mobility coefficient follows the idea of Grün and Rumpf [108] that requires the mesh to be composed of right-angled elements for $d = 2, 3$. Later, in Section 3.4 we prove well-posedness of this numerical scheme and give stability bounds. These results allow us to pass to the limit $\varepsilon \to 0$ in the regularized finite element problem and prove the existence of nonnegative global solutions of the non-regularized problem. We also prove the convergence of the discrete solutions to the weak solutions of the continuous relaxed Cahn-Hilliard model in dimension one. Then, aiming at designing a fully practical structure-preserving numerical scheme and based on the idea of the piecewise reconstruction of the mobility used in section 3.4, we propose a new adaptation of the upwind method inside the finite element method in Section 3.5. This allows us to design a new nonlinear semi-implicit scheme for the regularized problem. Then, we prove well-posedness and stability for this upwind scheme (i.e. we show the non-negativity preserving property and energy stability). We are able to prove for this numerical scheme the convergence in dimension one to the continuous regularized-relaxed model. Section 3.6 is devoted to the description of an efficient linear semi-implicit upwind scheme. Existence and non-negativity of a global solution is given using again the upwind approximation of the mobility coefficient. Finally, in Section 3.7, we present numerical simulations using our linear semi-implicit upwind scheme in one and two dimensions that are in good agreement with previous numerical results obtained for the degenerate Cahn-Hilliard equation with single-well logarithmic potential.

3.2 Notations

We first set up the notations we will use in the numerical discretization and recall some well-known properties we employ in the analysis of the scheme.

Geometric and functional setting. Let $\Omega \subset \mathbb{R}^d$ with $d = 1, 2, 3$ be a polyhedral domain. We indicate the usual Lebesgue and Sobolev spaces by respectively $L^p(\Omega)$, $W^{m,p}(\Omega)$ with $H^m(\Omega) := W^{m,2}(\Omega)$, where $1 \leq p \leq +\infty$ and $m \in \mathbb{N}$. We denote the corresponding norms by $||\cdot||_{m,p,\Omega}$, $||\cdot||_{m,\Omega}$ and semi-norms by $|\cdot|_{m,p,\Omega}$, $|\cdot|_{m,\Omega}$. The standard L^2 inner product will be denoted by $(\cdot,\cdot)_\Omega$ and the duality pairing between $(H^1(\Omega))'$ and $H^1(\Omega)$ by $<\cdot,\cdot>_\Omega$.
Let $\mathcal{T}^h$, $h > 0$ be a conformal mesh on the domain Ω which is defined by N_{el} disjoint piecewise linear mesh elements, denoted by $K \in \mathcal{T}^h$, such that $\overline{\Omega} = \bigcup_{K\in\mathcal{T}^h} \overline{K}$. The elements are triangles for $d = 2$ and tetrahedra for $d = 3$. We let $h := \max_K h_K$ refers to the level of refinement of the mesh, where $h_K := \text{diam}(K)$ for $K \in \mathcal{T}^h$. We define by κ_K the minimal perpendicular length of K and $\kappa_h = \min_{K\in\mathcal{T}^h} \kappa_K$. We assume that the mesh is quasi-uniform, *i.e.*, it is shape-regular and there exists a constant $C > 0$ such that

$$h_K \geq Ch, \quad \forall K \in \mathcal{T}^h.$$

Moreover, we assume that the mesh is acute, *i.e.*, for $d = 2$ the angles of the triangles do not exceed $\frac{\pi}{2}$ and for $d = 3$ the angle between two faces of the same tetrahedron do not exceed $\frac{\pi}{2}$. We define the set of nodal points $J = \{x_j\}_{j=1,\dots,N_h}$ of cardinality $N_h := \mathrm{card}(J)$ and we assume that each x_j is a vertex of a simplex $K \in \mathcal{T}^h$. Furthermore, in Section 3.4, we need to add an additional assumption about the mesh: the elements are right-angled.

We also define the barycentric dual mesh associated to $\mathcal{T}^h$. On every element $K \in \mathcal{T}^h$, the barycentric coordinates of an arbitrary point $x \in K$ are defined by the real numbers λ_i with $i = 1, \dots, n_K$ such that

$$\sum_{i=1}^{n_K} \lambda_i = 1, \quad \text{and} \quad x = \sum_{i=1}^{n_K} \lambda_i P_i,$$

where n_K is the number of nodes of the element K. We define the barycentric subdomains associated to the vertex $P_i \in K_k$ (where K_k refers to the k-th element of $\mathcal{T}^h$ and $k = 1, \dots, N_{\mathrm{el}}$), by

$$D_i^k := \bigcap_{\substack{j=1 \\ j \neq i}}^{n_K} \{x; x \in K_k \text{ and } \lambda_j(x) \leq \lambda_i(x)\}.$$

Therefore, for every node of the mesh $\mathcal{T}^h$, we define the barycentric cell of the dual mesh

$$D_i := \bigcup_k \{D_i^k;\ K_k \in \mathcal{T}^h \text{ such that } x_i \in K^k\}.$$

Therefore, we define these cells for all nodes of the mesh to define the barycentric dual mesh.

We introduce the set of piecewise linear functions $\chi_j \in C(\overline{\Omega})$ associated with the nodal point $x_j \in J$, that satisfies $\chi_j(x_i) = \delta_{ij}$, where δ_{ij} is the Kronecker's delta function. We introduce the $\mathbb{P}^1$ conformal finite element space V^h associated with $\mathcal{T}^h$, where $\mathbb{P}^1(K)$ denotes the space of polynomials of order 1 on K:

$$V^h := \left\{\chi \in C(\overline{\Omega}) : \chi\big|_K \in \mathbb{P}^1(K), \quad \forall K \in \mathcal{T}^h\right\} \subset H^1(\Omega).$$

Furthermore, we let K^h be the subset containing the nonnegative elements of V^h, namely

$$K^h := \{\chi \in V^h : \chi \geqslant 0 \quad \text{in } \Omega\}.$$

We denote by $\pi^h \colon C(\overline{\Omega}) \to V^h$ the Lagrange interpolation operator corresponding to the discretized domain $\mathcal{T}^h$, defined as

$$\pi^h f(x) = \sum_{j=1}^{N_h} f(x_j)\, \chi_j(x), \quad f \in C(\overline{\Omega}).$$

We also use the lumped space $\hat{V}_h$ defined by

$$\hat{V}_h := \{\hat{\chi} : \text{ piecewise constant over barycentric domains i.e. } \hat{\chi}(x) = \hat{\chi}(x_i),\ \forall x \in D_i\}.$$

Defining $\hat{\chi}_i \in L^\infty(\Omega)$ the characteristic function of the barycentric domain D_i associated with each node x_i (for $i = 1, \dots, N_h$) of the mesh, we easily see that $\{\hat{\chi}_j\}_{j=1,\dots,N_h}$ forms a basis of $\hat{V}_h$. Adding the property $\hat{\chi}_i(x_j) = \delta_{ij}$, we see that the two basis $\{\hat{\chi}_j\}_{j=1,\dots,N_h}$ and $\{\chi_j\}_{j=1,\dots,N_h}$ are associative i.e. $\chi(x_i) = \hat{\chi}(x_i)$ for all $x_i \in J$. Therefore, we also define the lumped operator

$\hat{\pi}^h : C(\overline{\Omega}) \to \hat{V}^h$ by

$$\hat{\pi}^h f(x) = \sum_{j=1}^{N_h} f(x_j)\, \hat{\chi}_j(x), \quad f \in C(\overline{\Omega}).$$

On $C(\overline{\Omega})$ we define the approximate scalar product as

$$(f,g)^h := \int_\Omega \pi^h \left(f(x)\, g(x)\right) \mathrm{d}x, \quad f,g \in C(\overline{\Omega}).$$

Furthermore, since $\forall f,g \in C(\overline{\Omega})$, we have

$$\begin{aligned}
\int_\Omega \pi^h \left(f(x)\, g(x)\right) \mathrm{d}x &= \sum_{K\in\mathcal{T}^h} \int_K \pi^h \left(f(x)\, g(x)\right) \mathrm{d}x, \\
&= \frac{1}{d+1} \sum_{K\in\mathcal{T}^h} |K| \sum_{x_i \in K} f(x_i)\, g(x_i), \\
&= \int_\Omega \hat{\pi}^h \left(f(x)\, g(x)\right) \mathrm{d}x.
\end{aligned}$$

where $x_i \in K$ denotes the vertices of the element K. We denote the corresponding discrete semi-norm as $|\cdot|_h = \left[(\cdot,\cdot)^h\right]^{\frac{1}{2}}$.

Continuous and discrete functionals. We denote by $P_h : L^2(\Omega) \to V^h$ the L^2 projection operator and by $\hat{P}_h : L^2(\Omega) \to V^h$ its lumped version, defined by

$$\begin{aligned}
(P_h v, \chi) &= (v,\chi) \quad \forall v \in L^2(\Omega) \text{ and } \forall \chi \in V^h, \\
\left(\hat{P}_h v, \chi\right)^h &= (v,\chi) \quad \forall v \in L^2(\Omega) \text{ and } \forall \chi \in V^h.
\end{aligned}$$

Furthermore, we introduce the inverse Laplacian operator $\mathcal{G} : \mathcal{F} \to S$ as an application from $\mathcal{F} = \{v \in \left(H^1(\Omega)\right)' :< v, 1 >= 0\}$ to $S = \{v \in H^1(\Omega) : (v,1) = 0\}$, that satisfies

$$(\nabla \mathcal{G} v, \nabla \eta) =< v, \eta > \quad \forall \eta \in H^1(\Omega). \tag{3.8}$$

The well-posedness of (3.8) follows immediately from the Lax-Milgram theorem and the Poincaré inequality. Therefore, a norm on $\mathcal{F}$ can be defined via

$$\|v\|_{\mathcal{F}} := |\mathcal{G} v|_1 \equiv \langle v, \mathcal{G} v\rangle^{\frac{1}{2}} \quad \forall v \in \mathcal{F}.$$

The discrete counterpart of $\mathcal{G}$ is denoted by $\hat{\mathcal{G}}^h : \mathcal{F}^h \to S^h$ and satisfies

$$\left(\nabla \hat{\mathcal{G}}^h v, \nabla \chi\right) = (v,\chi)^h \quad \forall \chi \in V^h,$$

where $S^h := \{v^h \in V^h : \left(v^h, 1\right) = 0\}$ and $\mathcal{F}^h := \{v \in C(\overline{\Omega}) : (v,1)^h = 0\}$.

Inequalities. We summarize important inequalities that will be used later on for the analysis of the numerical schemes. We start by recalling the well-known Sobolev interpolation result.

Letting $p \in [1, \infty]$, $m \geq 1$,

$$r \in \begin{cases} [p, \infty] & \text{if } m - \frac{d}{p} > 0, \\ [p, \infty) & \text{if } m - \frac{d}{p} = 0, \\ [p, -\frac{d}{m-(d/p)}] & \text{otherwise,} \end{cases}$$

and $\mu = \frac{d}{m}\left(\frac{1}{p} - \frac{1}{r}\right)$, there exists a constant $C = C(\Omega, p, r, m) > 0$ such that

$$\|v\|_{0,r} \leq C \|v\|_{0,p}^{1-\mu} \|v\|_{m,p}^{\mu} \quad \forall v \in W^{m,p}(\Omega). \tag{3.9}$$

Moreover, we will use the following inequalities (see, *e.g.*, [168]):

$$|\chi|_{m,p_2} \leq C h^{-d\left(\frac{1}{p_1} - \frac{1}{p_2}\right)} |\chi|_{m,p_1} \quad \forall \chi \in V^h, 1 \leq p_1 \leq p_2 \leq +\infty, m = 0, 1; \tag{3.10}$$

$$\|\chi\|_0^2 \leq (\chi, \chi)^h \leq (d+2) \|\chi\|_0^2 \quad \forall \chi \in V^h. \tag{3.11}$$

Concerning the interpolation operator, the following results hold [43]:

$$\lim_{h \to 0} \left\| v - \pi^h(v) \right\|_{0,\infty} = 0 \quad \forall v \in C(\overline{\Omega}), \tag{3.12}$$

and we have [192],

$$\left| (v, \eta)^h - (v, \eta) \right| \leq C h^2 \|\nabla v\|_0 \|\nabla \eta\|_0, \quad v, \eta \in V^h. \tag{3.13}$$

Furthermore, if $d = 1$ (see for example [192]),

$$\left| v - \pi^h(v) \right|_{m,p} \leq C h^{1-m} |v|_{1,p} \quad \forall v \in W^{1,p}(\Omega), \quad m = 0, 1, \quad 1 \leq p < +\infty; \tag{3.14}$$

$$\left\| v - \pi^h(v) \right\|_{L^\infty(\Omega)} + h \left| v - \pi^h(v) \right|_{1,\infty} \leq C h^2 |v|_{1,\infty} \quad \forall v \in H^1(\Omega), \tag{3.15}$$

$$\left| (v, \eta)^h - (v, \eta) \right| \leq C \left(\left| v - \pi^h v \right|_0 + h |v|_0 \right) \|\eta\|_1, \quad \text{for} \quad v \in C\left(\overline{\Omega}\right), \eta \in H^1(\Omega). \tag{3.16}$$

For the L^2 projection operator the following inequalities hold

$$|v - P_h v|_0 + h |v - P_h v|_1 \leq C h^m \|v\|_m \quad v \in H^m(\Omega), \quad m = 1, 2, \tag{3.17}$$

and for the lumped version

$$\left| v - \hat{P}_h v \right|_0 + h \left| v - \hat{P}_h v \right|_1 \leq C h \|v\|_1 \quad v \in H^1(\Omega).$$

Finally, the discrete inverse laplacian operator satisfies

$$(v, \chi)^h \equiv \left(\nabla \hat{\mathcal{G}}^h v, \nabla \chi \right) \leq \left| \hat{\mathcal{G}}^h v \right|_1 |\chi|_1 \quad \forall v \in \mathcal{F}^h, \chi \in V^h. \tag{3.18}$$

Finite element matrices. We define M and Q respectively the mass and stiffness matrix. M_l is the lumped mass matrix, that is the diagonal matrix where each coefficient is the sum of the associated row of M

$$M_{ij} = \int_\Omega \chi_i \chi_j \, \mathrm{d}x, \quad \text{for } i, j = 1, \ldots, N_h,$$

$$Q_{ij} = \int_\Omega \nabla\chi_i \nabla\chi_j \,\mathrm{d}x, \qquad \text{for } i,j = 1,\dots,N_h,$$

$$M_{l,ii} := \sum_{j=1}^{N_h} M_{ij}, \qquad \text{for } i,j = 1,\dots,N_h.$$

3.3 Definition of the regularized problem

As for the continuous case (see [165]), we use a regularization of the model. The resulting problem is easier to analyze since the singularity contained in the potential ψ_+ is smoothed out and the mobility is no longer degenerate.

Regularization of the mobility and potential. We define the regularized problem similarly to [165]. We consider a small positive parameter $0 < \varepsilon \ll 1$ and define the regularized mobility

$$b_\varepsilon(n) := \begin{cases} b(1-\varepsilon) & \text{for } n \geq 1-\varepsilon, \\ b(\varepsilon) & \text{for } n \leq \varepsilon, \\ b(n) & \text{otherwise.} \end{cases} \tag{3.19}$$

Therefore, there are two positive constants b_1 and B_1 such that

$$b_1 < b_\varepsilon(n) < B_1, \quad \forall n \in \mathbb{R}, \tag{3.20}$$

and the regularized mobility satisfies

$$b_\varepsilon \in C(\mathbb{R}, \mathbb{R}^+). \tag{3.21}$$

To define the regularized potential, we smooth out the singularity contained in ψ_+ and located at $n = 1$, see (3.6). We define for all $n \in \mathbb{R}$

$$\psi''_{+,\varepsilon}(n) := \begin{cases} \psi''_+(1-\varepsilon) & \text{for } n \geq 1-\varepsilon, \\ \psi''_+(\varepsilon) & \text{for } n \leq \varepsilon, \\ \psi''_+(n) & \text{otherwise.} \end{cases} \tag{3.22}$$

We can easily prove that for all $n \geq 1-\varepsilon$, $\psi_{+,\varepsilon}$ is bounded from below. Therefore, it exists a positive finite constant C_1 such that

$$\psi_{+,\varepsilon}(n) > \frac{1-n^\star}{2\varepsilon^2}\left([n-1]_+\right)^2 - C_1, \qquad \forall n \in \mathbb{R}, \tag{3.23}$$

where $[\cdot]_+ = \max\{\cdot, 0\}$.
Then, denoting by $\overline{\psi}_-(n)$ the extension of ψ_- given in (3.6) on all $\mathbb{R}$, it exists a constant $C_2 > 0$ such that the regularized potential $\psi_\varepsilon(n) = \psi_{+,\varepsilon}(n) + \overline{\psi}_-(n)$ satisfies

$$\psi_\varepsilon(n) \geq \frac{1-n^\star}{2\varepsilon^2}\left([n-1]_+\right)^2 - C_2, \qquad \forall n \in \mathbb{R}. \tag{3.24}$$

Altogether, we obtain

$$\psi_\varepsilon \in C^2(\mathbb{R}, \mathbb{R}). \tag{3.25}$$

Regularized problem. The regularized-relaxed degenerate Cahn-Hilliard model reads

$$\begin{cases} \partial_t n_\varepsilon = \nabla \cdot \left[b_\varepsilon(n_\varepsilon) \nabla (\varphi_\varepsilon + \psi'_{+,\varepsilon}(n_\varepsilon)) \right], \\ -\sigma \Delta \varphi_\varepsilon + \varphi_\varepsilon = -\gamma \Delta n_\varepsilon + \overline{\psi}'_-(n_\varepsilon - \frac{\sigma}{\gamma}\varphi_\varepsilon), \end{cases} \quad t > 0,\ x \in \Omega, \tag{3.26}$$

with zero-flux boundary conditions

$$\frac{\partial (n_\varepsilon - \frac{\sigma}{\gamma}\varphi_\varepsilon)}{\partial \nu} = b_\varepsilon(n_\varepsilon) \frac{\partial \big(\varphi_\varepsilon + \psi'_{+,\varepsilon}(n_\varepsilon)\big)}{\partial \nu} = 0 \qquad \text{on } \partial\Omega \times (0, +\infty). \tag{3.27}$$

3.4 Nonlinear semi-implicit scheme

In this section, we assume that the mesh is composed of right-angled elements for $d = 2, 3$.

3.4.1 Description of the nonlinear numerical scheme.

The finite element problem associated with (3.26) is: For each $k = 0, \dots, N_T - 1$, find $\{n_{h,\varepsilon}^{k+1}, \varphi_{h,\varepsilon}^{k+1}\}$ in $V^h \times V^h$ such that $\forall \chi \in V^h$ we have

$$\begin{cases} \left(\dfrac{n_{h,\varepsilon}^{k+1} - n_h^k}{\Delta t}, \chi \right)^h + \left(\tilde{M}_\varepsilon(n_{h,\varepsilon}^{k+1}) \nabla \left(\varphi_{h,\varepsilon}^{k+1} + \pi^h(\psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})) \right), \nabla \chi \right) = 0, & \text{(3.28a)} \\ \sigma \left(\nabla \varphi_{h,\varepsilon}^{k+1}, \nabla \chi \right) + \left(\varphi_{h,\varepsilon}^{k+1}, \chi \right)^h = \gamma \left(\nabla n_{h,\varepsilon}^{k+1}, \nabla \chi \right) + \left(\overline{\psi}'_-(n_h^k - \dfrac{\sigma}{\gamma}\varphi_h^k), \chi \right)^h, & \text{(3.28b)} \end{cases}$$

where $n_{h,\varepsilon}^{k+1} = \sum_{i=1,\dots,N_h} n_\varepsilon(x_i, t^{k+1})\chi_i$, and $\varphi_{h,\varepsilon}^{k+1} = \sum_{i=1,\dots,N_h} \varphi_\varepsilon(x_i, t^{k+1})\chi_i$.

The initial condition $n_h^0 \in V^h$ is given by

$$\begin{cases} n_h^0 = \pi^h n^0(x), & \text{if } d = 1, \\ n_h^0 = \hat{P}_h n^0(x), & \text{if } d = 2, 3, \end{cases}$$

and φ_h^0 is the solution of

$$\sigma \left(\nabla \varphi_h^0, \nabla \chi \right) + \left(\varphi_h^0, \chi \right)^h = \gamma \left(\nabla n_h^0, \nabla \chi \right) + \left(\psi'_-(n_h^0 - \frac{\sigma}{\gamma}\varphi_h^0), \chi \right)^h, \quad \forall \chi \in V^h. \tag{3.29}$$

The well-posedness of equation (3.29) follows the Lax-Milgram theorem.

For $k = 0, \dots, N_T - 1$, let $\underline{n}^k$ and $\underline{\varphi}^k$ be the vectors

$$\underline{n}^k := [n_1^k, \dots, n_{N_h}^k]^T, \quad \underline{\varphi}^k := [\varphi_1^k, \dots, \varphi_{N_h}^k]^T.$$

We can then rewrite the problem in its matrix form

$$M_l \underline{n}^{k+1} = M_l \underline{n}^k - \Delta t U \underline{\psi}'_+ - \Delta t U \underline{\varphi}^{k+1}, \tag{3.30}$$

$$(\sigma Q + M_l)\, \underline{\varphi}^{k+1} = \gamma Q \underline{n}^{k+1} + M_l \underline{\psi}'_-, \tag{3.31}$$

where $\underline{\psi}'_+$ and $\underline{\psi}'_-$ are the two vectors containing the values at the nodes of the functionals

$$\left(\underline{\psi}'_+\right)_i = \frac{1-n^\star}{1-\underline{n}_i^{k+1}} - \left(\underline{n}_i^{k+1}\right)^2 \quad i=1,\dots,N_h,$$
$$\left(\underline{\psi}'_-\right)_i = -(1-n^\star)\left(\underline{n}_i^k - \frac{\sigma}{\gamma}\underline{\varphi}_i^k\right) - (1-n^\star) \quad i=1,\dots,N_h.$$

We denote by U the finite element matrix corresponding to the term $\left(\tilde{M}_\varepsilon(n_h^{k+1})\nabla\cdot,\nabla\cdot\right)$.

We define the $d\times d$ matrix $\tilde{M}_\varepsilon$ that approximates the continuous mobility. To do so, we define the entropy functional $\phi_\varepsilon:\mathbb{R}\to\mathbb{R}^+$ such that

$$\phi''_\varepsilon(s) = \frac{1}{b_\varepsilon(s)}, \quad \forall s\in\mathbb{R}. \tag{3.32}$$

Then, we define the following properties for the mobility $\tilde{M}_\varepsilon : V^h \to \otimes_{k=1}^{|\mathcal{T}^h|}\mathbb{R}^{d\times d}$:

i) $\tilde{M}_\varepsilon : V^h \to \otimes_{k=1}^{|\mathcal{T}^h|}\mathbb{R}^{d\times d}$ is continuous;

ii) $\tilde{M}_\varepsilon(s)\Big|_K = b_\varepsilon(s)\mathrm{I}_d$ if s is constant on the element $K\in\mathcal{T}^h$;

iii) $\tilde{M}_\varepsilon^T(s_h)\nabla\pi^h(\phi'_\varepsilon(s_h)) = \nabla s_h$, $\forall s_h\in V^h$;

iv) on each element $K\in\mathcal{T}^h$, the matrix $\tilde{M}_\varepsilon(s)\Big|_K$ is symmetric and positive semidefinite.

For each element $K\in\mathcal{T}^h$, $\tilde{M}_\varepsilon(s_h)$ is a $d\times d$ matrix. For $d\le 3$, we consider a reference element $\hat{K} = \hat{K}_{(\alpha_1,\dots,\alpha_d)}$ where the corners of this element are defined by

$$\hat{x}_0 = 0 \text{ and } \hat{x}_i = \alpha_i e_i, \quad \forall i=1,\dots,d \text{ and } \alpha_i\in\mathbb{R}.$$

Here, e_i denotes the i-th unit vector. In the following, we denote by p_0 and p_i the nodes in the physical space that correspond to $\hat{x}_0$ and $\hat{x}_i$. For each element K of the triangulation $\mathcal{T}^h$ it exists an orthogonal matrix A such that the linear affine mapping $H:\hat{K}\to K$ defined by $\hat{x}\to x = p_0 + A\hat{x}$ is a bijection. Then, we set for any element $K\in\mathcal{T}^h$ and $s_h\in V^h$

$$\tilde{M}_\varepsilon(s_h)\Big|_K = A\hat{M}_\varepsilon(\hat{s}_h)A^{-1}, \quad \text{with} \quad \hat{s}_h(\hat{x}) = s_h(H\hat{x}).$$

Then, we define $\hat{M}_\varepsilon$ on the element K to be a diagonal matrix with

$$\hat{M}_{\varepsilon,ii}(\hat{s}_h) = \frac{\hat{s}_h(\hat{x}_i)-\hat{s}_h(\hat{x}_0)}{\phi'_\varepsilon(\hat{s}_h(\hat{p}_i))-\phi'_\varepsilon(\hat{s}_h(\hat{p}_0))} = \frac{s_h(p_i)-s_h(p_0)}{\phi'_\varepsilon(s_h(p_i))-\phi'_\varepsilon(s_h(p_0))} = \frac{1}{\phi''_\varepsilon(s_h(\eta))}, \tag{3.33}$$

for some η between p_0 and p_i, if $s_h(p_i)\neq s_h(p_0)$. Then, if $s_h(p_i)=s_h(p_0)$ we set

$$\hat{M}_{\varepsilon,ii}(\hat{s}_h) = \frac{1}{\phi''_\varepsilon(s_h(p_0))}. \tag{3.34}$$

Altogether, we obtain the following result

Proposition 17 *The previous definitions of the mobility $\tilde{M}_\varepsilon$ and entropy ϕ_ε satisfies the axioms i)–iv).*

Proof. From the definition of the mobility $\tilde{M}_\varepsilon$, if the quantity $n_{h,\varepsilon}^{k+1} \in V^h$ is constant on the element, we have $M_\varepsilon(n_{h,\varepsilon}^{k+1})\Big|_{\hat{K}} = b_\varepsilon(n_{h,\varepsilon}^{k+1})\mathrm{I_d}$. Then, using the definition of the second derivative of the entropy, we have

$$\begin{aligned}\phi_\varepsilon'(n_{h,\varepsilon}^{k+1}(x_i)) - \phi_\varepsilon'(n_{h,\varepsilon}^{k+1}(x_0)) &= \int_{n_{h,\varepsilon}^{k+1}(x_0)}^{n_{h,\varepsilon}^{k+1}(x_i)} \phi_\varepsilon''(s)\ \mathrm{d}s \\ &= \int_{n_{h,\varepsilon}^{k+1}(x_0)}^{n_{h,\varepsilon}^{k+1}(x_i)} \frac{1}{b_\varepsilon(s)}\ \mathrm{d}s.\end{aligned}$$

Therefore, from the definition of $\hat{M}_\varepsilon$, we have that

$$\hat{M}_\varepsilon^T \nabla_{\hat{x}} \pi^h(\phi_\varepsilon'(n_{h,\varepsilon}^{k+1})) = \nabla_{\hat{x}} n_{h,\varepsilon}^{k+1},$$

where $\nabla_{\hat{x}}$ denotes the gradient on the reference element $\hat{K}$.

Therefore, defining $M_\varepsilon = A\hat{M}_\varepsilon A^{-1}$, the axioms ii), iii) and iv) (because A is orthogonal) are satisfied on any $K \in \mathcal{T}^h$. To conclude, since the same procedure can be applied for any element of $\mathcal{T}^h$, the axiom i) is satisfied.

□

3.4.2 Well-posedness of the regularized problem and stability bounds

Existence of discrete solutions and energy dissipation.

Theorem 18 *Let $d \leq 3$, the system (3.28a)–(3.28b) with an initial condition satisfying $n_h^0 \in K^h$, has a solution $\{n_{h,\varepsilon}^{k+1}, \varphi_{h,\varepsilon}^{k+1}\} \in V^h \times V^h$.*

Furthermore, the solutions satisfy the inequality

$$E(n_{h,\varepsilon}^{k+1}, \varphi_{h,\varepsilon}^{k+1}) + \Delta t \sum_{k=0}^{N_T-1} \int_\Omega \tilde{M}_\varepsilon(n_{h,\varepsilon}^{k+1}) \left|\nabla\left(\varphi_{h,\varepsilon}^{k+1} + \pi^h\left(\psi_{+,\varepsilon}'(n_{h,\varepsilon}^{k+1})\right)\right)\right|^2 dx \leqslant E(n_h^0, \varphi_h^0)\,, \tag{3.35}$$

where

$$\begin{aligned}E(n_{h,\varepsilon}^{k+1}, \varphi_{h,\varepsilon}^{k+1}) := \frac{\gamma}{2}\left|n_{h,\varepsilon}^{k+1} - \frac{\sigma}{\gamma}\varphi_{h,\varepsilon}^{k+1}\right|_1^2 + \frac{\sigma}{2\gamma}\left\|\varphi_{h,\varepsilon}^{k+1}\right\|_0^2 \\ + \left(\psi_{+,\varepsilon}(n_{h,\varepsilon}^{k+1}) + \overline{\psi}_-\left(n_{h,\varepsilon}^{k+1} - \frac{\sigma}{\gamma}\varphi_{h,\varepsilon}^{k+1}\right), 1\right)^h.\end{aligned} \tag{3.36}$$

Proof. *Step 1. Existence of global in time solutions.* To prove existence of discrete solutions to (3.28a) and (3.28b), we use the Brouwer fixed point theorem and we adapt the proof from [108]. In the following, all vector quantities are denoted using the convention

$$s_h \in V^h \Rightarrow \underline{s} = (s_h(x_1), \ldots, s_h(x_{N_h})).$$

We define $w_h^k = n_h^k - \alpha$, where $\alpha := \frac{1}{|\Omega|}\int_\Omega n_h^0$. The system of equations (3.28a)–(3.28b) becomes

$$\begin{aligned}\left(w_{h,\varepsilon}^{k+1} - w_h^k, \chi\right)^h &= -\Delta t\left(\tilde{M}_\varepsilon(w_{h,\varepsilon}^{k+1}+\alpha)\nabla\left(\varphi_{h,\varepsilon}^{k+1} + \pi^h\left(\psi'_{+,\varepsilon}(w_{h,\varepsilon}^{k+1}+\alpha)\right)\right), \nabla\chi\right),\\ \left(\varphi_{h,\varepsilon}^{k+1}, \chi\right)^h &= \gamma\left(\nabla\left(w_{h,\varepsilon}^{k+1} - \frac{\sigma}{\gamma}\varphi_{h,\varepsilon}^{k+1}\right), \nabla\chi\right) + \left(\overline{\psi}'_-(w_h^k + \alpha - \frac{\sigma}{\gamma}\varphi_h^k), \chi\right)^h.\end{aligned} \tag{3.37}$$

Given $w_h^k \in V^h$ with $-\alpha \le w_h^k \le 1-\alpha$, we want to prove the existence of a solution $w_{h,\varepsilon}^{k+1} \in V^h$ such that $R(\underline{w}^{k+1}) = \underline{w}^{k+1}$ with

$$\begin{aligned}-R(w) = F(w) =& \left[\left(\Delta t M_l^{-1} U_\varepsilon(\underline{w}+\underline{\alpha})\left(M_l + \sigma Q\right)^{-1}(\gamma Q)\right)\right]\underline{w}\\ &+ \Delta t M_l^{-1} U_\varepsilon(\underline{w}+\underline{\alpha})\left(\psi'_{+,\varepsilon}(\underline{w}+\underline{\alpha})\right)\\ &+ \Delta t M_l^{-1} U_\varepsilon(\underline{w}+\underline{\alpha})\left(M_l+\sigma Q\right)^{-1}\underline{r}^k - \underline{w}^k,\end{aligned}$$

where $\underline{r}^k$ is the vector associated with $M_l\left(\overline{\psi}'_-(\underline{w}_h^k + \underline{\alpha} - \frac{\sigma}{\gamma}\underline{\varphi}_h^k)\right)$. To apply the Brouwer fixed point theorem, we want to prove that $F : \tilde{K}^h \to \tilde{K}^h$ is a Lipschitz continuous mapping on

$$\tilde{K}^h = \{w \in V^h | M_l \underline{w} \cdot (1, \dots, 1) = 0\},$$

which is a convex subspace of V^h. This constraint on the space $\tilde{K}^h$ reflects the conservation of the mass. Let us compute

$$\begin{aligned}F(\underline{w}) - F(\underline{w}^k) =& \left(\Delta t M_l^{-1}\left(U_\varepsilon(\underline{w}+\underline{\alpha}) - U_\varepsilon(\underline{w}^k+\underline{\alpha})\right)\left(M_l+\sigma Q\right)^{-1}(\gamma Q)\right)(\underline{w}-\underline{w}^k)\\ &+\Delta t M_l^{-1}\left(U_\varepsilon(\underline{w}+\underline{\alpha}) - U_\varepsilon(\underline{w}^k+\underline{\alpha})\right)\left(\psi'_{+,\varepsilon}(\underline{w}+\underline{\alpha}) - \psi'_{+,\varepsilon}(\underline{w}^k+\underline{\alpha})\right)\\ &+\Delta t M_l^{-1}\left(U_\varepsilon(\underline{w}+\underline{\alpha}) - U_\varepsilon(\underline{w}^k+\underline{\alpha})\right)\left(M_l+\sigma Q\right)^{-1}\left(\underline{r}^k - \underline{r}^{k-1}\right)\\ &+(\underline{w}^k - \underline{w}^{k-1}).\end{aligned}$$

Using the continuity of $\overline{\psi}'_-$ and the fact that $w^k \in \tilde{K}^h$ is bounded, we have that it exists a positive constant C such that

$$\begin{aligned}\left\|F(\underline{w}) - F(\underline{w}^k)\right\| \le& \Delta t\left\|M_l^{-1}\left(U_\varepsilon(\underline{w}+\underline{\alpha}) - U_\varepsilon(\underline{w}^k+\underline{\alpha})\right)\left(M_l+\sigma Q\right)^{-1}(\gamma Q)\right\|\left\|\underline{w}-\underline{w}^k\right\|\\ &+\Delta t\left\|M_l^{-1}\left(U_\varepsilon(\underline{w}+\underline{\alpha}) - U_\varepsilon(\underline{w}^k+\underline{\alpha})\right)\right\|\left\|\psi'_{+,\varepsilon}(\underline{w}+\underline{\alpha}) - \psi'_{+,\varepsilon}(\underline{w}^k+\underline{\alpha})\right\| + C.\end{aligned}$$

Then, from the continuity of $\psi'_{+,\varepsilon}$, we know that

$$\left\|\psi'_{+,\varepsilon}(\underline{w}+\underline{\alpha}) - \psi'_{+,\varepsilon}(\underline{w}^k+\underline{\alpha})\right\| \le C\left\|\underline{w}-\underline{w}^k\right\|.$$

Since $M_l + \sigma Q$ is a M-matrix, the norm of its inverse is bounded from Varah's bound [197]. Altogether and using the fact that the mobility $\tilde{M}_\varepsilon(s_h)$ is bounded for all $s_h \in V^h$, we obtain

$$\left\|F(\underline{w}) - F(\underline{w}^k)\right\| \le C(\Delta t, h)\left\|\underline{w}-\underline{w}^k\right\|,$$

which proves that the mapping F is Lipschitz continuous, and hence R is also Lipschitz continuous on the convex set $\tilde{K}^h$. Therefore, applying the Brouwer fixed point theorem, the mapping R admits a fixed point. Consequently, the system (3.28a) and (3.28b) admits a solution $n_{h,\varepsilon}^{k+1} = w_{h,\varepsilon}^{k+1} + \alpha$ globally in time. Then, since $\varphi_{h,\varepsilon}^{k+1}$ is uniquely defined by $n_{h,\varepsilon}^{k+1}, \varphi_h^k, n_h^k$ from

the equation of (3.28b), it exists a pair of functions $\{n_{h,\varepsilon}^{k+1}, \varphi_{h,\varepsilon}^{k+1}\} \in V^h \times V^h$ solution of the problem.

Step 2. Energy estimate. We prove that the discrete solutions of the regularized problem satisfy the energy inequality (3.35). Let us start by rewriting equation (3.28a) for $j = 1, \dots, N_h$,

$$\left(n_{h,\varepsilon}^{k+1} - n_h^k, \chi_j\right)^h = -\Delta t \left(\tilde{M}_\varepsilon(n_{h,\varepsilon}^{k+1}) \nabla \left(\varphi_{h,\varepsilon}^{k+1} + \pi^h \left(\psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right)\right), \nabla \chi_j\right). \tag{3.38}$$

Using the definition of the lumped scalar product, we have

$$\begin{aligned}\left(n_{h,\varepsilon}^{k+1} - n_h^k, \chi_j\right)^h &= \sum_{x_i \in J_h} (1, \chi_i) \left(n_{h,\varepsilon}^{k+1} - n_h^k\right)(x_i) \chi_j(x_i), \\ &= (1, \chi_j) \left(n_{h,\varepsilon}^{k+1} - n_h^k\right)(x_j).\end{aligned}$$

Multiplying the previous equation by $\left(\varphi_{h,\varepsilon}^{k+1} + \psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right)(x_j)$ and summing over $x_j \in J$, we obtain

$$\begin{aligned}\sum_{x_j \in J} (1, \chi_j) \left(n_{h,\varepsilon}^{k+1} - n_h^k\right)(x_j) \left(\varphi_{h,\varepsilon}^{k+1} + \psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right)(x_j) \\ = \left(n_{h,\varepsilon}^{k+1} - n_h^k, \varphi_{h,\varepsilon}^{k+1} + \psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right)^h.\end{aligned}$$

Repeating the same operations on the right-hand side of (3.38) gives

$$\begin{aligned}-\sum_{x_j \in J_h} \Delta t \left(\tilde{M}_\varepsilon(n_{h,\varepsilon}^{k+1}) \nabla \left(\varphi_{h,\varepsilon}^{k+1} + \pi^h \left(\psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right)\right), \nabla \chi_j\right) \left(\varphi_{h,\varepsilon}^{k+1} + \psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right)(x_j) \\ = -\Delta t \int_\Omega \tilde{M}_\varepsilon(n_{h,\varepsilon}^{k+1}) \left|\nabla \left(\varphi_{h,\varepsilon}^{k+1} + \pi^h \left(\psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right)\right)\right|^2.\end{aligned}$$

We now focus on the term $\left(n_{h,\varepsilon}^{k+1} - n_h^k, \psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right)^h$. We observe that, for a convex function g, the following property holds

$$g(y) - g(x) \leqslant g'(y)(y - x).$$

Thus, since $\psi_{+,\varepsilon}(\cdot)$ is the convex part of the potential, we have

$$\begin{aligned}\left(n_{h,\varepsilon}^{k+1} - n_h^k, \psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right)^h &= \sum_{x_j \in J} (1, \chi_j) \left(n_{h,\varepsilon}^{k+1} - n_h^k\right)(x_j) \psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1}(x_j)) \\ &\geq \sum_{x_j \in J} (1, \chi_j) \left(\psi_{+,\varepsilon}(n_{h,\varepsilon}^{k+1}(x_j)) - \psi_{+,\varepsilon}(n_h^k(x_j))\right) \\ &= \left(\psi_{+,\varepsilon}(n_{h,\varepsilon}^{k+1}) - \psi_{+,\varepsilon}(n_h^k), 1\right)^h.\end{aligned} \tag{3.39}$$

We need now to bound from below the term $\left(n_{h,\varepsilon}^{k+1} - n_h^k, \varphi_{h,\varepsilon}^{k+1}\right)^h$. First, using the symmetry of

the scalar product, we observe that

$$\left(n_{h,\varepsilon}^{k+1}-n_h^k,\varphi_{h,\varepsilon}^{k+1}\right)^h=\left(\varphi_{h,\varepsilon}^{k+1},\left(n_{h,\varepsilon}^{k+1}-\frac{\sigma}{\gamma}\varphi_{h,\varepsilon}^{k+1}\right)-\left(n_h^k-\frac{\sigma}{\gamma}\varphi_h^k\right)\right)^h + \frac{\sigma}{\gamma}\left(\varphi_{h,\varepsilon}^{k+1}-\varphi_h^k,\varphi_{h,\varepsilon}^{k+1}\right)^h.$$

Taking $\chi=\left(n_{h,\varepsilon}^{k+1}-\frac{\sigma}{\gamma}\varphi_{h,\varepsilon}^{k+1}\right)-\left(n_h^k-\frac{\sigma}{\gamma}\varphi_h^k\right)$, which is an admissible test function since all the quantities are in the correct space S^h, we obtain

$$\begin{aligned}&\left(\varphi_{h,\varepsilon}^{k+1},\left(n_{h,\varepsilon}^{k+1}-\frac{\sigma}{\gamma}\varphi_{h,\varepsilon}^{k+1}\right)-\left(n_h^k-\frac{\sigma}{\gamma}\varphi_h^k\right)\right)^h\\&=\gamma\left(\nabla\left(n_{h,\varepsilon}^{k+1}-\frac{\sigma}{\gamma}\varphi_{h,\varepsilon}^{k+1}\right),\nabla\left(\left(n_{h,\varepsilon}^{k+1}-\frac{\sigma}{\gamma}\varphi_{h,\varepsilon}^{k+1}\right)-\left(n_h^k-\frac{\sigma}{\gamma}\varphi_h^k\right)\right)\right)\\&+\left(\overline{\psi}_-'\left(n_h^k-\frac{\sigma}{\gamma}\varphi_h^k\right),\left(n_{h,\varepsilon}^{k+1}-\frac{\sigma}{\gamma}\varphi_{h,\varepsilon}^{k+1}\right)-\left(n_h^k-\frac{\sigma}{\gamma}\varphi_h^k\right)\right)^h.\end{aligned}$$

Using the elementary property

$$a(a-b)\geqslant\frac{1}{2}\left(a^2-b^2\right), \tag{3.40}$$

we obtain

$$\begin{aligned}&\gamma\left(\nabla\left(n_{h,\varepsilon}^{k+1}-\frac{\sigma}{\gamma}\varphi_{h,\varepsilon}^{k+1}\right),\nabla\left(\left(n_{h,\varepsilon}^{k+1}-\frac{\sigma}{\gamma}\varphi_{h,\varepsilon}^{k+1}\right)-\left(n_h^k-\frac{\sigma}{\gamma}\varphi_h^k\right)\right)\right)\\&\geq\frac{\gamma}{2}\left(\left|n_{h,\varepsilon}^{k+1}-\frac{\sigma}{\gamma}\varphi_{h,\varepsilon}^{k+1}\right|_1^2-\left|n_h^k-\frac{\sigma}{\gamma}\varphi_h^k\right|_1^2\right),\end{aligned} \tag{3.41}$$

and

$$\frac{\sigma}{\gamma}\left(\varphi_{h,\varepsilon}^{k+1}-\varphi_h^k,\varphi_{h,\varepsilon}^{k+1}\right)^h\geq\frac{\sigma}{2\gamma}\left(\left|\varphi_{h,\varepsilon}^{k+1}\right|_h^2-\left|\varphi_h^k\right|_h^2\right). \tag{3.42}$$

Then, noting that $\overline{\psi}_-(\cdot)$ is concave, we obtain the inequality

$$\begin{aligned}&\left(\overline{\psi}_-'\left(n_h^k-\frac{\sigma}{\gamma}\varphi_h^k\right),\left(n_{h,\varepsilon}^{k+1}-\frac{\sigma}{\gamma}\varphi_{h,\varepsilon}^{k+1}\right)-\left(n_h^k-\frac{\sigma}{\gamma}\varphi_h^k\right)\right)^h\\&\geq\left(\overline{\psi}_-\left(n_{h,\varepsilon}^{k+1}-\frac{\sigma}{\gamma}\varphi_{h,\varepsilon}^{k+1}\right)-\overline{\psi}_-\left(n_h^k-\frac{\sigma}{\gamma}\varphi_h^k\right),1\right)^h.\end{aligned} \tag{3.43}$$

Combining the inequalities (3.39), (3.41), (3.42), (3.43) and using the definition of the discrete energy (3.36) yields to

$$E_h(n_{h,\varepsilon}^{k+1},\varphi_{h,\varepsilon}^{k+1})+\Delta t\int_\Omega\tilde{M}_\varepsilon(n_{h,\varepsilon}^{k+1})\left|\nabla\left(\varphi_{h,\varepsilon}^{k+1}+\pi^h\left(\psi_+'(n_{h,\varepsilon}^{k+1})\right)\right)\right|^2\mathrm{d}x\leqslant E(n_h^k,\varphi_h^k).$$

Then summing the previous equation from $k=0\to N_T-1$, we obtain (3.35).

□

Discrete entropy inequality. We managed to obtain bounds for important quantities but we still need to find an inequality for $\left|\varphi_{h,\varepsilon}^{k+1}\right|_1$ and $\left|n_{h,\varepsilon}^{k+1}\right|_1$. To tackle this issue, we use the previously

defined entropy-mobility pair.

Therefore, we can state the following result.

Theorem 19 (Entropy estimate) *The solutions* $\{n_{h,\varepsilon}^{k+1}, \varphi_{h,\varepsilon}^{k+1}\}$ *of the system* (3.28a)–(3.28b) *with the mobility defined by* (3.33)–(3.34) *and the entropy defined by* (3.32)*, satisfy the inequality*

$$\left(\phi_\varepsilon(n_{h,\varepsilon}^{k+1}), 1\right)^h + \min\left(\psi_{+,\varepsilon}''\right)\left|n_{h,\varepsilon}^{k+1}\right|_1^2 + \frac{\sigma}{\gamma}\left|\varphi_{h,\varepsilon}^{k+1}\right|_1^2 \leq \left(\phi_\varepsilon(n_h^k), 1\right)^h + C. \tag{3.44}$$

Proof. We use as a test function $\chi = \pi^h(\phi_\varepsilon'(n_{h,\varepsilon}^{k+1}))$ in (3.28a), we have

$$\begin{aligned}&\left(n_{h,\varepsilon}^{k+1} - n_h^k, \pi^h(\phi'(n_{h,\varepsilon}^{k+1}))\right)^h \\ &= -\left(\tilde{M}_\varepsilon(n_{h,\varepsilon}^{k+1})\nabla\left(\varphi_{h,\varepsilon}^{k+1} + \pi^h\left(\psi_{+,\varepsilon}'(n_{h,\varepsilon}^{k+1})\right)\right), \nabla\pi^h(\phi'(n_{h,\varepsilon}^{k+1}))\right).\end{aligned} \tag{3.45}$$

Let us focus on the left-hand side. Using the convexity of ϕ_ε, we obtain

$$\left(n_{h,\varepsilon}^{k+1} - n_h^k, \pi^h(\phi'(n_{h,\varepsilon}^{k+1}))\right)^h = \left(n_{h,\varepsilon}^{k+1} - n_h^k, \phi_\varepsilon'(n_{h,\varepsilon}^{k+1})\right)^h \geq \left(\phi_\varepsilon(n_{h,\varepsilon}^{k+1}) - \phi_\varepsilon(n_h^k), 1\right)^h.$$

Then, let us rewrite the right-hand side of (3.45) using the axiom iii) by

$$\begin{aligned}-\left(\tilde{M}_\varepsilon(n_{h,\varepsilon}^{k+1})\nabla\left(\varphi_{h,\varepsilon}^{k+1} + \pi^h\left(\psi_{+,\varepsilon}'(n_{h,\varepsilon}^{k+1})\right)\right), \nabla\pi^h(\phi_\varepsilon'(n_{h,\varepsilon}^{k+1}))\right) = \\ -\left(\nabla\left(\varphi_{h,\varepsilon}^{k+1} + \pi^h\left(\psi_{+,\varepsilon}'(n_{h,\varepsilon}^{k+1})\right)\right), \nabla n_{h,\varepsilon}^{k+1}\right).\end{aligned}$$

Combining the definition of the Lagrange interpolation operator and the regularized potential, we have

$$\pi^h\left(\psi_{+,\varepsilon}'(n_{h,\varepsilon}^{k+1})\right) = \sum_{x_j \in J_h} \psi_{+,\varepsilon}'(n_{h,\varepsilon}^{k+1}(x_j))\chi_j.$$

Consequently, we have

$$\pi^h(\psi_{+,\varepsilon}'(n_{h,\varepsilon}^{k+1})) = f(n_{h,\varepsilon}^{k+1}), \quad \text{for } \varepsilon \leq n_{h,\varepsilon}^{k+1} \leq 1-\varepsilon,$$

where $f(n_{h,\varepsilon}^{k+1})$ is a strictly positive function which increases monotonically with $n_{h,\varepsilon}^{k+1}$. We also have

$$\pi^h\left(\psi_{+,\varepsilon}'(n_{h,\varepsilon}^{k+1})\right) = \sum_{x_j \in J} \psi_{+,\varepsilon}''(n_{h,\varepsilon}^{k+1}(x_j))n_{h,\varepsilon}^{k+1} \text{ for } n_{h,\varepsilon}^{k+1}(x_j) \leq \varepsilon \text{ and } n_{h,\varepsilon}^{k+1}(x_j) \geq 1-\varepsilon.$$

Therefore, we obtain that $\nabla\pi^h(n_{h,\varepsilon}^{k+1})$ behaves as $\nabla n_{h,\varepsilon}^{k+1}$. Altogether, we obtain

$$\left(\nabla\pi^h\left(\psi_{+,\varepsilon}'(n_{h,\varepsilon}^{k+1})\right), \nabla n_{h,\varepsilon}^{k+1}\right) \geq \min\left(f'(n_{h,\varepsilon}^{k+1})\right)\left|n_{h,\varepsilon}^{k+1}\right|_1^2,$$

and we know that $\min\left(f'(n_{h,\varepsilon}^{k+1})\right)$ is positive. The last term to handle is $\left(\nabla\varphi_{h,\varepsilon}^{k+1}, \nabla n_{h,\varepsilon}^{k+1}\right)$.

Using the equation (3.28b) with $\chi = \varphi_{h,\varepsilon}^{k+1}$, we obtain

$$\left(\nabla n_{h,\varepsilon}^{k+1}, \nabla \varphi_{h,\varepsilon}^{k+1}\right) = \frac{1}{\gamma}\left[\sigma \left|\varphi_{h,\varepsilon}^{k+1}\right|_1^2 + \left|\varphi_{h,\varepsilon}^{k+1}\right|_h^2 - \left(\overline{\psi}'_- \left(n_h^k - \frac{\sigma}{\gamma}\varphi_h^k\right), \varphi_{h,\varepsilon}^{k+1}\right)^h\right]$$
$$\leq \frac{\sigma}{\gamma}\left|\varphi_{h,\varepsilon}^{k+1}\right|_1^2 + \left(\frac{1}{2\gamma}\right)\left|\varphi_{h,\varepsilon}^{k+1}\right|_h^2,$$

where the last inequality is obtained using Young's inequality and the energy inequality. Altogether, we obtain the inequality (3.44).

□

Convergence $\varepsilon \to 0$, $h > 0$

Theorem 20 *For every sequence $\varepsilon \to 0$, we can extract subsequences such that we have*

$$n_{h,\varepsilon'}^{k+1} \to n_h^{k+1} \quad \text{and} \quad \nabla n_{h,\varepsilon'}^{k+1} \to \nabla n_h^{k+1}, \tag{3.46}$$

where $n_h^{k+1} \in K^h$. Similarly, we also have

$$\varphi_{h,\varepsilon'}^{k+1} \to \varphi_h^{k+1} \quad \text{and} \quad \nabla \varphi_{h,\varepsilon'}^{k+1} \to \nabla \varphi_h^{k+1}, \tag{3.47}$$

where $\varphi_h^{k+1} \in V^h$.

Proof. From the inequalities (3.35) and (3.44), we have that both $\left|n_{h,\varepsilon}^{k+1}\right|_1^2$ and $\left|\varphi_{h,\varepsilon}^{k+1}\right|_1^2$ are bounded. Then, from the Poincaré-Wirtinger inequality, we obtain the following convergences of subsequences

$$n_{h,\varepsilon'}^{k+1} \to n_h^{k+1} \quad \text{and} \quad \nabla n_{h,\varepsilon'}^{k+1} \to \nabla n_h^{k+1} \quad \text{in} \quad V^h,$$
$$\varphi_{h,\varepsilon'}^{k+1} \to \varphi_h^{k+1} \quad \text{and} \quad \nabla \varphi_{h,\varepsilon'}^{k+1} \to \nabla \varphi_h^{k+1} \quad \text{in} \quad V^h.$$

We now prove that the limit belongs to K^h. Using (3.24) and (3.35), we obtain

$$\left(\left[n_{h,\varepsilon}^{k+1} - 1\right]^2, 1\right) \leq C\varepsilon^2,$$

from which we can conclude using (3.11) and (3.10)

$$\left\|\left[n_{h,\varepsilon}^{k+1} - 1\right]\right\|_{0,\infty}^2 \leq Ch^{-d/2}\varepsilon.$$

Next, we want to prove the existence of a small parameter ε_0 such that $n^{k+1} < 1$ for each $\varepsilon \leq \varepsilon_0$. This result can be obtained from the fact that $\psi_{+,\varepsilon}(n_{h,\varepsilon}^{k+1}) \geq 0$ for $n_{h,\varepsilon}^{k+1} \geq 0$ and from (3.35) leading to

$$\left(\psi_{+,\varepsilon}(n_{h,\varepsilon}^{k+1}), \psi_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right)^h \leq C\left(\psi_{+,\varepsilon}(n_{h,\varepsilon}^{k+1}), 1\right)^h \leq C.$$

Again from (3.11) and (3.10) we have the bound

$$\left\|\psi_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right\|_{0,\infty} \leq Ch^{-d/2}, \tag{3.48}$$

which is independent on ε. Hence, we conclude that $n_{h,\varepsilon}^{k+1} < 1$ because if it was not the case

(3.48) will yield to a contradiction due to the logarithmic term in $\psi_+(\cdot)$. Consequently,

$$n_h^{k+1} < 1.$$

To prove that $0 \leq n_{h,\varepsilon}^{k+1}$ in Ω we use the discrete analogue of the argument from the continuous setting [165]. For $\alpha > 0$, we define the following set

$$V_\alpha^\varepsilon = \{x_i \text{ nodal point } | -n_{h,\varepsilon}^{k+1}(x_i) \geq \alpha\}.$$

Thus, we have that for $A > 0$, it exists a small ε_0 such that

$$\phi_\varepsilon''(n_{h,\varepsilon}^{k+1}) \geq 2A, \quad \forall n_{h,\varepsilon}^{k+1} \leq 0, \qquad \forall \varepsilon \leq \varepsilon_0.$$

Integrating the previous inequality twice, we obtain

$$\phi_\varepsilon(n_{h,\varepsilon}^{k+1}) \geq A(n_{h,\varepsilon}^{k+1})^2.$$

Since the discrete entropy is bounded uniformly in ε, we have

$$A\alpha^2 \left|V_\alpha^\varepsilon\right| \leq \left(\phi_\varepsilon(n_{h,\varepsilon}^{k+1}), 1\right)^h \leq C.$$

Therefore, the measure of the set V_α^ε is bounded and using the strong convergence $n_{h,\varepsilon}^{k+1} \to n_h^{k+1}$ (which is obtained from the fact that both entropy and energy bounds are uniform in ε) as well as Fatou's lemma, we obtain

$$\left|\{x_i \text{ nodal point } | -n_h^{k+1}(x_i) \geq \alpha\}\right| \leq \frac{C}{A\alpha^2} \quad \forall A > 0.$$

Consequently, we obtain that $n_h^{k+1} \geq 0$. Altogether, we have proved that the limit solution belongs to

$$\{n_h^{k+1}, \varphi_h^{k+1}\} \in K^h \times V^h.$$

□

3.4.3 Well-posedness of the non regularized problem and stability

The results we established for the regularized problem allow us to study the degenerate system by passing to the limit as epsilon vanishes. Therefore, we set the non-regularized finite element problem:

For each $k = 0, \ldots, N_T - 1$, find $\{n_h^{k+1}, \varphi_h^{k+1}\}$ in $V^h \times V^h$ such that $\forall \chi \in V^h$ we have

$$\begin{cases} \left(\dfrac{n_h^{k+1} - n_h^k}{\Delta t}, \chi\right)^h + \left(\tilde{M}(n_h^{k+1})\nabla\left(\varphi_h^{k+1} + \pi^h(\psi_+'(n_h^{k+1}))\right), \nabla\chi\right) = 0\,, & \text{(3.49a)} \\ \sigma\left(\nabla\varphi_h^{k+1}, \nabla\chi\right) + \left(\varphi_h^{k+1}, \chi\right)^h = \gamma\left(\nabla n_h^{k+1}, \nabla\chi\right) + \left(\overline{\psi}_-'(n_h^k - \dfrac{\sigma}{\gamma}\varphi_h^k), \chi\right)^h\,, & \text{(3.49b)} \end{cases}$$

where $n_h^{k+1} = \sum_{i=1,\ldots,N_h} n(x_i, t^{k+1})\chi_i$, and $\varphi_h^{k+1} = \sum_{i=1,\ldots,N_h} \varphi_(x_i, t^{k+1})\chi_i$.

Then, we establish the following theorem

Theorem 21 (Well-posedness and stability bound) *Let $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$, the system (3.49a)–(3.49b) with initial condition $n_h^0 \in K^h$ admits a solution $\{n_h^{k+1}, \varphi_h^{k+1}\} \in K^h \times V^h$.*

Furthermore, the solution $\{n_h^{k+1}, \varphi_h^{k+1}\}$ of problem (3.49a)–(3.49b) satisfies

$$\begin{aligned}\max_{k=0\to N_T-1} \left(\left\|n_h^{k+1}\right\|_1^2 + \left\|\varphi_h^{k+1}\right\|_1^2 \right) + (\Delta t)^2 \sum_{k=0}^{N_T-1} \left(\left\| \frac{n_h^{k+1}-n_h^k}{\Delta t} \right\|_1^2 + \left\| \frac{\varphi_h^{k+1}-\varphi_h^k}{\Delta t} \right\|_1^2 \right) \\ + \sum_{k=0}^{N_T-1} \Delta t \left| \left(\tilde{M}(n_h^{k+1}) \right)^{\frac{1}{2}} \nabla \left(\varphi_h^{k+1} + \pi^h \left(\psi_+'(n_h^{k+1}) \right) \right) \right|_0^2 \\ + \sum_{k=0}^{N_T-1} \Delta t \left(\tilde{B}_{max} \right)^{-1} \left| \hat{\mathcal{G}}^h \left[\frac{n_h^{k+1}-n_h^k}{\Delta t} \right] \right|_1^2 \leq C(n^0),\end{aligned} \tag{3.50}$$

where $\tilde{B}_{max} = \left\| \tilde{M}(n_h^{k+1}) \right\|_\infty$.

Proof. *Step 1. Well-posedness.* Going back to the regularized problem (3.28a)–(3.28b), we can pass to the limit using the strong convergences (3.46)–(3.47) to obtain

$$\lim_{\varepsilon\to 0} \left(\frac{n_{h,\varepsilon}^{k+1}-n_h^k}{\Delta t}, \chi \right)^h = \frac{n_h^{k+1}-n_h^k}{\Delta t}, \tag{3.51}$$

$$\begin{aligned}\lim_{\varepsilon\to 0} \left(\tilde{M}_\varepsilon(n_{h,\varepsilon}^{k+1}) \nabla \left(\varphi_{h,\varepsilon}^{k+1} + \pi^h \left(\psi_{+,\varepsilon}'(n_{h,\varepsilon}^{k+1}) \right) \right), \nabla\chi \right) \\ = \left(\tilde{M}(n_h^{k+1}) \nabla \left(\varphi_h^{k+1} + \pi^h \left(\psi_+'(n_h^{k+1}) \right) \right), \nabla\chi \right),\end{aligned} \tag{3.52}$$

$$\lim_{\varepsilon\to 0} \left(\nabla\varphi_{h,\varepsilon}^{k+1}, \nabla\chi \right) + \left(\varphi_{h,\varepsilon}^{k+1}, \chi \right)^h = \left(\nabla\varphi_h^{k+1}, \nabla\chi \right) + \left(\varphi_h^{k+1}, \chi \right)^h, \tag{3.53}$$

$$\begin{aligned}\lim_{\varepsilon\to 0} \left(\nabla n_{h,\varepsilon}^{k+1}, \nabla\chi \right) + \left(\overline{\psi}_-'(n_h^k - \frac{\sigma}{\gamma}\varphi_{h,\varepsilon}^k), \chi \right)^h \\ = \left(\nabla n_h^{k+1}, \nabla\chi \right) + \left(\psi_-'(n_h^k - \frac{\sigma}{\gamma}\varphi_h^k), \chi \right)^h.\end{aligned} \tag{3.54}$$

Step 2. Stability bound. First of all, we state the energy inequality for the non-regularized problem using the fact that (3.35) is independent of ε. Hence, we have

$$\begin{aligned}\frac{\gamma}{2} \left| n_h^{k+1} - \frac{\sigma}{\gamma}\varphi_h^{k+1} \right|_1^2 + \frac{\sigma}{2\gamma} \left\|\varphi_h^{k+1}\right\|_0^2 + \left(\psi_+(n_h^{k+1}) + \psi_- \left(n_h^{k+1} - \frac{\sigma}{\gamma}\varphi_h^{k+1} \right), 1 \right)^h \\ + \Delta t \sum_{k=0}^{N_T-1} \int_\Omega \tilde{M}(n_h^{k+1}) \left| \nabla \left(\varphi_h^{k+1} + \pi^h \left(\psi_+'(n_h^{k+1}) \right) \right) \right|^2 \mathrm{d}x \leq C,\end{aligned} \tag{3.55}$$

where we assumed that the initial energy has a finite value. Similarly, we have the entropy inequality

$$\sum_{k=0}^{N_T-1} \left| n_h^{k+1} \right|_1^2 + \frac{\sigma}{\gamma} \left| \varphi_h^{k+1} \right|_1^2 \leq C. \tag{3.56}$$

Let us now use $\chi = 2\Delta t n_h^{k+1}$ in (3.49a) to obtain

$$\left(n_h^{k+1} - n_h^k, 2n_h^{k+1}\right)^h = -2\Delta t \int_\Omega \tilde{M}(n_h^{k+1}) \nabla \left(\varphi_h^{k+1} + \pi^h \left(\psi_+'(n_h^{k+1})\right)\right) \nabla n_h^{k+1} \, \mathrm{d}x.$$

Using the relation $2a(a-b) = a^2 - b^2 + (a-b)^2$, the fact that $n_h^{k+1} \in K^h$ and the Cauchy-Schwarz inequality, we have

$$\begin{aligned} &\left|n_h^{k+1}\right|_h^2 - \left|n_h^k\right|_h^2 + (\Delta t)^2 \left|\frac{n_h^{k+1} - n_h^k}{\Delta t}\right|_h^2 \\ &\qquad \leq 2 \left((\Delta t)^2 \tilde{B}_{\max} \int_\Omega \tilde{M}(n_h^{k+1}) \left|\nabla \left(\varphi_h^{k+1} + \pi^h \left(\psi_+'(n_h^{k+1})\right)\right)\right|^2\right)^{\frac{1}{2}} \left|n_h^{k+1}\right|_1 \\ &\qquad \leq C, \end{aligned} \tag{3.57}$$

where the upper bound is obtained from the energy inequality (3.55) and the entropy inequality (3.56). Moreover, taking $\chi = n_h^{k+1} - n_h^k$ in (3.49a) and using the identity $2a(a-b) = a^2 - b^2 + (a-b)^2$, we have

$$\begin{aligned} &\frac{\gamma}{2}\left(\left|n_h^{k+1}\right|_1^2 - \left|n_h^k\right|_1^2 + \left|n_h^{k+1} - n_h^k\right|_1^2\right) \\ &\quad = \sigma\left(\varphi_h^{k+1}, n_h^{k+1} - n_h^k\right) + \left(\varphi_h^{k+1}, n_h^{k+1} - n_h^k\right)^h - \left(\psi_-'(n_h^k - \frac{\sigma}{\gamma}\varphi_h^k), n_h^{k+1} - n_h^k\right)^h. \end{aligned}$$

However, we know by definition that

$$\begin{aligned} &-\left(\psi_-'(n_h^k - \frac{\sigma}{\gamma}\varphi_h^k), n_h^{k+1} - n_h^k\right)^h \\ &\qquad = (1 - n^\star)\left(\left(n_h^k - \frac{\sigma}{\gamma}\varphi_h^k, n_h^{k+1} - n_h^k\right)^h + \left(1, n_h^{k+1} - n_h^k\right)^h\right), \end{aligned}$$

which is bounded from above using Cauchy-Schwarz inequality, (3.57) and (3.56). Therefore, using Young's inequality and (3.11), we obtain

$$\begin{aligned} \frac{(\gamma - \sigma)}{2}\left|n_h^{k+1} - n_h^k\right|_1^2 \leq \frac{\gamma}{2}&\left(\left|n_h^{k+1}\right|_1^2 - \left|n_h^k\right|_1^2\right) \\ &+ \frac{\sigma}{2}\left|\varphi_h^{k+1}\right|_1^2 + \frac{d+2}{2}\left(\left\|\varphi_h^{k+1}\right\|_0^2 + \left\|n_h^{k+1} - n_h^k\right\|_0^2\right) + C. \end{aligned}$$

Then, subtracting equation (3.49b) at the two times t^{k+1} and t^k, we get

$$\begin{aligned} \left(\nabla\left(\varphi_h^{k+1} - \varphi_h^k\right), \nabla\chi\right) + \left(\varphi_h^{k+1} - \varphi_h^k, \chi\right)^h &= \left(\nabla\left(n_h^{k+1} - n_h^k\right), \chi\right) \\ &+ \left(\psi_-'\left(n_h^k - \frac{\sigma}{\gamma}\varphi_h^k\right) - \psi_-'\left(n_h^{k-1} - \frac{\sigma}{\gamma}\varphi_h^{k-1}\right), \chi\right)^h. \end{aligned}$$

Then, taking $\chi = \varphi_h^{k+1} - \varphi_h^k$ and using Young's inequality, there are two positive constants κ_1

and κ_2 chosen such that $\sigma - \frac{\gamma\kappa_1}{2} > 0$, and $1 - \frac{\kappa_2}{2} > 0$.

$$\begin{aligned}\left(\sigma - \frac{\gamma^2\kappa_1}{2}\right)\left|\varphi_h^{k+1} - \varphi_h^k\right|_1^2 + \left(1 - \frac{\kappa_2}{2}\right)\left|\varphi_h^{k+1} - \varphi_h^k\right|_h^2 \\ = \frac{1}{2\kappa_1}\left|n_h^{k+1} - n_h^k\right|_1^2 + \frac{1}{2\kappa_2}\left|\psi_-'\left(n_h^k - \frac{\sigma}{\gamma}\varphi_h^k\right) - \psi_-'\left(n_h^{k-1} - \frac{\sigma}{\gamma}\varphi_h^{k-1}\right)\right|_h^2.\end{aligned} \tag{3.58}$$

Using the fact that $\psi_-'(\cdot)$ is linear and the $L^2(\Omega)$ norms of n_h^k, φ_h^k, n_h^{k-1} and φ_h^{k-1} are finite, together with the energy inequality (3.55), we obtain an upper bound for the right-hand side of (3.58). Then, using (3.11), we have the existence of a positive constant such that

$$\left\|\varphi_h^{k+1} - \varphi_h^k\right\|_1^2 \leq C.$$

Finally, taking $\chi = \hat{\mathcal{G}}^h\left[\frac{n_h^{k+1}-n_h^k}{\Delta t}\right]$ in (3.49a), we have

$$\left(\frac{n_h^{k+1} - n_h^k}{\Delta t}, \hat{\mathcal{G}}^h\left[\frac{n_h^{k+1} - n_h^k}{\Delta t}\right]\right)^h = \left|\hat{\mathcal{G}}^h\left[\frac{n_h^{k+1} - n_h^k}{\Delta t}\right]\right|_1^2,$$

and

$$\begin{aligned}\left|\hat{\mathcal{G}}^h\left[\frac{n_h^{k+1} - n_h^k}{\Delta t}\right]\right|_1^2 &= -\left(\tilde{M}(n_h^{k+1})\nabla\left(\varphi_h^{k+1} + \pi^h(\psi_+'(n_h^{k+1}))\right), \nabla\hat{\mathcal{G}}^h\left[\frac{n_h^{k+1} - n_h^k}{\Delta t}\right]\right) \\ &\leq \left|\tilde{M}(n_h^{k+1})\nabla\left(\varphi_h^{k+1} + \pi^h(\psi_+'(n_h^{k+1}))\right)\right|_0^2 \\ &\leq \tilde{B}_{\max}\left|\left(\tilde{M}(n_h^{k+1})\right)^{\frac{1}{2}}\nabla\left(\varphi_h^{k+1} + \pi^h(\psi_+'(n_h^{k+1}))\right)\right|_0^2.\end{aligned}$$

Altogether and summing from $k = 0 \to N_T - 1$, we obtain (3.50).

3.4.4 Convergence analysis

In order to study the convergence of the scheme as $h, \Delta t \to 0$, we follow [28] and define for $k = 0, \ldots, N_T - 1$

$$U_h(t,x) := \frac{t - t_k}{\Delta t} n_h^{k+1} + \frac{t_{k+1} - t}{\Delta t} n_h^k, \quad t \in (t_k, t_{k+1}],$$

and

$$U_h^+ := n_h^{k+1}, \quad U_h^- := n_h^k.$$

First we remark that, thanks to (3.50), $U_h \in L^2(0,T;H^1(\Omega))$. Moreover, simple calculations show that for $t \in (t_k, t_{k+1}]$

$$\frac{\partial U_h}{\partial t} = \frac{n_h^{k+1} - n_h^k}{\Delta t} \quad t \in (t_k, t_{k+1}], \quad k \geq 0,$$

and

$$U_h - U_h^+ = (t - t_{k+1})\frac{\partial U_h}{\partial t}, \quad \text{as well as} \quad U_h - U_h^- = (t - t_k)\frac{\partial U_{h,\varepsilon}}{\partial t} \quad t \in (t_k, t_{k+1}], \quad k \geq 0.$$

We also have the analogous definition for $W_{h,\varepsilon}$ which is

$$W_h(t,x) := \frac{t - t_k}{\Delta t}\varphi_h^{k+1} + \frac{t_{k+1} - t}{\Delta t}\varphi_h^k, \quad t \in (t_k, t_{k+1}],$$

and we also have

$$W_h^+ := \varphi_h^{k+1}, \quad W_h^- := \varphi_h^k.$$

We can state the following theorem:

Theorem 22 (Convergence) *Let $d = 1, 2, 3$ and $n^0 \in H^1(\Omega)$, with $0 \leq n^0 < 1$ a.e. Ω. We assume that $\{\mathcal{T}^h, n_h^0, \Delta t\}_{h>0}$ satisfy*

i) $n_h^0 \in V^h$ with $0 \leq n_h^0 < 1$, given by $n_h^0 = \pi^h(n^0)$ if $d = 1$, and $\hat{P}_h(n^0)$ if $d = 2, 3$.

ii) Let $\Omega \subset \mathbb{R}^d$ be a polyhedral domain and $\mathcal{T}^h$ a quasi-uniform acute mesh of it into N right-angled mesh elements.

Therefore, for $\Delta t, h \to 0$, it exists a subsequence of solutions $\{U_h, W_h\}$ and a pair of function

$$\{n, \varphi\} \in L^\infty\left(0, T; H^1(\Omega)\right) \cap C_{x,t}^{\frac{1}{2},\frac{1}{8}}(\overline{\Omega}_T) \cap H^1(0, T; (H^1(\Omega))') \times L^\infty\left(0, T; H^1(\Omega)\right) \quad \text{if} \quad d = 1,$$
$$\{n, \varphi\} \in L^\infty\left(0, T; H^1(\Omega)\right) \cap H^1(0, T; (H^1(\Omega))') \times L^\infty\left(0, T; H^1(\Omega)\right) \quad \text{if} \quad d = 2, 3,$$

with

$$0 \leq n < 1, \quad \text{a.e. in } \Omega_T,$$

such that

$$U_h, U_h^+, U_h^- \to n, \quad \text{uniformly on } \overline{\Omega}_T, \quad \text{if} \quad d = 1, \tag{3.59}$$
$$U_h, U_h^+, U_h^- \to n, \quad \text{strongly in } L^\infty\left(0, T; L^2(\Omega)\right), \quad \text{if} \quad d = 2, 3, \tag{3.60}$$
$$U_h, U_h^+, U_h^- \rightharpoonup n, \quad \text{weakly in } L^\infty\left(0, T; H^1(\Omega)\right), \tag{3.61}$$
$$\frac{\partial U_h}{\partial t} \rightharpoonup \frac{\partial n}{\partial t}, \quad \text{weakly in } L^2\left(0, T; \left(H^1(\Omega)\right)'\right), \tag{3.62}$$
$$W_h, W_h^+, W_h^- \rightharpoonup \varphi, \quad \text{weakly in } L^\infty\left(0, T; H^1(\Omega)\right). \tag{3.63}$$

Moreover, for $d = 1$, $\{n, \varphi\}$ is a solution of the relaxed-degenerate Cahn-Hilliard equation under the weak form

$$\begin{cases} \int_0^T < \chi, \partial_t n > & = -\int_{\Omega_T} b(n)\nabla\left(\varphi + \psi_+'(n)\right) \cdot \nabla\chi, \quad \forall \chi \in L^2(0, T; H^1(\Omega)), \\ \int_{\Omega_T} \varphi\chi & = \int_{\Omega_T} \gamma\nabla\left(n - \frac{\sigma}{\gamma}\varphi\right) \cdot \nabla\chi + \psi_-'\left(n - \frac{\sigma}{\gamma}\varphi\right)\chi, \quad \forall \chi \in L^2(0, T; H^1(\Omega)). \end{cases} \tag{3.64}$$

Proof. *Step 1. Weak and strong convergences.* From the inequality (3.50), we know that

$$\begin{aligned} &\|U_h\|^2_{L^\infty(0,T;H^1(\Omega))} + \Delta t \|U_h\|^2_{H^1(0,T;H^1(\Omega))} \\ &\qquad + \left\| \left[\tilde{M}(U_h)\right]^{1/2} \nabla\left(W_h + \pi^h\left(\psi_+'(U_h)\right)\right) \right\|^2_{L^2(\Omega_T)} \leq C, \end{aligned} \tag{3.65}$$

and

$$\|W_h\|^2_{L^\infty(0,T;H^1(\Omega))} + \Delta t \, \|W_h\|^2_{H^1(0,T;H^1(\Omega))} \leq C. \tag{3.66}$$

Since

$$U_h - U_h^{\pm} = (t - t^{\pm,k}) \frac{\partial U_h}{\partial t},$$

with $t^{+,k} = t^{k+1}$ and $t^{-,k} = t^k$, we have from (3.65),

$$\left\| U_h - U_h^{\pm} \right\|^2_{L^2(0,T;H^1(\Omega))} \leq \Delta t^2 \left\| \frac{\partial U_h}{\partial t} \right\|^2_{L^2(0,T;H^1(\Omega))} \leq C \Delta t. \tag{3.67}$$

Using (3.66), the same can be applied with $W_h, W_h^{\pm}$ to obtain

$$\left\| W_h - W_h^{\pm} \right\|^2_{L^2(0,T;H^1(\Omega))} \leq \Delta t^2 \left\| \frac{\partial W_h}{\partial t} \right\|^2_{L^2(0,T;H^1(\Omega))} \leq C \Delta t. \tag{3.68}$$

The weak convergences (3.61), (3.62), (3.63) are obtained from the use of the inequalities (3.65), (3.66), (3.67), (3.68), and standard compactness results. Then, the strong convergence (3.60) is obtained from (3.61), (3.62) and by application of the Lions-Aubin lemma. Let us show that the discrete solution U_h is Hölder continuous for $d = 1$. From (3.65) and by Sobolev embeddings, we have

$$|U_h(x_2, t) - U_h(x_1, t)| \leq C \, |x_2 - x_1|^{\frac{1}{2}} \quad \forall x_1, x_2 \in \overline{\Omega}, \forall t \geq 0. \tag{3.69}$$

Furthermore, from (3.9), (3.11), (3.18), (3.50) and (3.65), we get

$$\begin{aligned}
\|U_h(x,t_2) - U_h(x,t_1)\|_{0,\infty} &\leq C \, \|U_h(x,t_2) - U_h(x,t_1)\|_0^{\frac{1}{2}} \, \|U_h(x,t_2) - U_h(x,t_1)\|_1^{\frac{1}{2}} \\
&\leq C \left| \hat{\mathcal{G}}^h \left(U_h(x,t_2) - U_h(x,t_1) \right) \right|_1^{\frac{1}{4}} \|U_h(x,t_2) - U_h(x,t_1)\|_1^{\frac{3}{4}} \\
&\leq C \left| \hat{\mathcal{G}}^h \left[\int_{t_1}^{t_2} \frac{\partial U_h}{\partial t}(x,t) \, \mathrm{d}t \right] \right|_1^{\frac{1}{4}} \left(2 \, \|U_h\|_{L^\infty(0,T;H^1(\Omega))} \right) \\
&\leq C \left| \int_{t_1}^{t_2} \hat{\mathcal{G}}^h \frac{\partial U_h}{\partial t}(x,t) \, \mathrm{d}t \right|_1^{\frac{1}{4}} \\
&\leq C (t_2 - t_1)^{\frac{1}{8}} \left(\int_{t_1}^{t_2} \left| \hat{\mathcal{G}}^h \frac{\partial U_h}{\partial t}(x,t) \right|_1^2 \mathrm{d}t \right)^{\frac{1}{8}} \\
&\leq C (t_2 - t_1)^{\frac{1}{8}} \quad \forall x \in \Omega, \, t_1, t_2 \geq 0.
\end{aligned} \tag{3.70}$$

Then, from (3.65), (3.69) and (3.70), we obtain that the $C^{\frac{1}{2},\frac{1}{8}}_{x,t}(\overline{\Omega}_T)$ norm is bounded independently of Δt, h and T. Therefore, every sequence $\{U_h\}_h$ is uniformly bounded and equicontinuous on $\overline{\Omega}_T$ with $T > 0$. Hence, from the use of the Arzelà-Ascoli theorem we obtain the convergence (3.59).

Step 2. Limiting equation for $d = 1$. We start by considering $\eta \in L^2(0,T;H^1(\Omega))$, and we

take $\chi = \pi^h \eta$ in (3.28b) to obtain

$$\int_0^T \left[\sigma\left(\nabla W_h^+, \nabla \pi^h \eta\right) + \left(W_h^+, \pi^h \eta\right)^h\right] \mathrm{d}t = \int_0^T \left[\gamma\left(\nabla U_h^+, \nabla \pi^h \eta\right) + \left(\psi'_-(U_h^- - \frac{\sigma}{\gamma} W_h^-), \pi^h \eta\right)^h\right] \mathrm{d}t.$$

Then, combining the weak convergences (3.61) and (3.63) with the properties (3.13), and (3.15), we can pass to the limit in the left-hand side and the first term of the right-hand side. Since ψ'_- is a linear functional, the two convergences (3.61) and (3.63) together with (3.13) are sufficient to pass to the limit.

Secondly, we show that for all $\eta \in L^2(0,T;H^1(\Omega))$, we have

$$\int_\Omega \left(\tilde{M}(U_h^+) - b(U_h^+) I_d\right) \nabla \left(W_h^+ + \pi^h \left(\psi'_{+,\varepsilon}(U_h^+)\right)\right) \nabla \pi^h \eta \, \mathrm{d}x \to 0, \quad \text{as} \quad h \to 0. \tag{3.71}$$

Since, $\tilde{M}$ is a piecewise constant approximation of b on all $K \in \mathcal{T}^h$, we know that $\tilde{M}(\cdot) \to b(\cdot) I_d$ uniformly and we obtain the previous convergence using a generalized version of the Lebesgue dominated convergence theorem. Therefore, combining the previous convergence (3.71), the strong convergence (3.60), the weak convergence (3.63), and (3.15), we obtain

$$\int_\Omega \tilde{M}(U_h^+) \nabla W_h^+ \nabla \pi^h \eta \, \mathrm{d}x \to \int_\Omega b(n) \nabla \varphi \nabla \eta \, \mathrm{d}x.$$

Then, from combining the convergence (3.71), the strong convergence (3.60), and (3.15), we obtain

$$\int_\Omega \tilde{M}(U_h^+) \nabla \pi^h \psi'_+(U_h^+) \nabla \pi^h \eta \, \mathrm{d}x \to \int_\Omega b(n) \psi''_+(n) \nabla n \nabla \eta \, \mathrm{d}x.$$

From the previous results, we show that if we take $\chi = \pi^h \eta$ with $\eta \in H^1\left(0,T;H^1(\Omega)\right)$ in (3.28a), the right-hand side converges to the right-hand side of the first equation of the expected limit system.

Finally, we focus on the left hand side of (3.28a) with $\chi = \pi^h \eta$ and $\eta \in H^1\left(0,T;H^1(\Omega)\right)$. From integration by parts we obtain

$$\int_0^T \left(\frac{\partial U_h}{\partial t}, \pi^h \eta\right)^h \mathrm{d}t = -\int_0^T \left(U_h, \frac{\partial \pi^h \eta}{\partial t}\right)^h \mathrm{d}t + \left(U_h(T), \pi^h \eta(T)\right)^h - \left(U_h(0), \pi^h \eta(0)\right)^h. \tag{3.72}$$

Furthermore, since $\left\{\tilde{M}(U_h^+) \nabla \left(W_h^+ + \pi^h \psi'_+(U_h^+)\right)\right\}_{h>0}$ is uniformly bounded in $L^2(\Omega_T)$, we know that $u \in H^1\left(0,T;(H^1(\Omega))'\right)$ Therefore, from (3.13), the weak convergence (3.61), equation (3.72) converges to the left-hand side of the first equation of the limit system.

Altogether, we recover the limit system (3.64).

□

3.5 Non-linear semi-implicit multi-dimensional upwind numerical scheme

As we have seen in the previous section, to preserve the non-negativity of the discrete solutions, a particular approximation of the mobility function is needed. Based upon the results obtained on finite volume schemes for nonlinear parabolic models, we propose an adaptation of the upwind method within the finite element method.

Upwind approximation of mobility. We approximate the continuous mobility $b(u_h^{k+1})$ by a piecewise continuous function $\tilde{B}(n_h^{k+1})$. This latter is constant on specific subdomains that we define for each element. We consider for each element $K \in \mathcal{T}^h$, the decomposition of K in $(d+1)$ subdomains defined by

$$\tilde{D}_{ij}^K = \{x \in K | \lambda_i, \lambda_j \geq \lambda_k, \quad k \neq i, j\},$$

for $i = 1, 2, 3$, $j = 2, 3$, and $i \neq j$. Setting $\xi_i^{k+1} := \left(\varphi_h^{k+1} + \psi'_+(n_h^{k+1})\right)(x_i)$, we define on each of the subdomains $\tilde{D}_{ij}^K$ (for each $K \in \mathcal{T}^h$)

$$B_{ij}^k := \begin{cases} n_i^k(1-n_j^k)^2, & \text{if } \xi_i^{k+1} - \xi_j^{k+1} > 0, \\ n_j^k(1-n_i^k)^2, & \text{otherwise.} \end{cases} \tag{3.73}$$

Each entries of the finite element matrix is approximated by

$$U_{ij} = \int_\Omega \tilde{B}(u_h^k) \nabla\chi_i \nabla\chi_j \, \mathrm{d}x \approx B_{ij}^k Q_{ij}, \quad \forall i, j = 1, \ldots, N_h.$$

The previous approximation is equivalent to a one point quadrature where the quadrature node is chosen to be part of the subdomain $\tilde{D}_{ij}$. In our case, the quadrature error of the multi-dimensional upwind method is

$$\int_\Omega \left| B_{ij}^k - \tilde{B}(u^k) \right| \mathrm{d}x \leq C \sum_{\substack{x_j \in \bigcup K \in \mathcal{T}^h \\ x_i, x_j \in K \\ x_j \neq x_i}} \left| u_h^k(x_i) - u_h^k(x_j) \right|. \tag{3.74}$$

The computation of the mobility coefficient (3.73) is similar to the one used in [12] for the one-dimensional finite volume discretization of the Keller-Segel system. Indeed, in one dimension, our method reduces exactly to a finite volume method. However, in higher dimensions the computation presents some differences. Definition (3.73) in the finite element context is also close in spirit to the one proposed by Baba and Tabata in [19], where the authors used barycentric coordinates to define the basis functions.
Our method is well suited for an assembling procedure and, as a result, is simpler to implement in already existing finite element software since it requires only the adaptation of the calculation of a non-constant matrix. This method can also be adapted for the simulation of other advection-diffusion equations to preserve the nonnegativity of solutions.

Non-linear semi-implicit upwind discretization for the regularized problem. As in the previous section, we start by describing the finite element problem associated to the regularized problem (3.26).

For each $k = 0, \dots, N_T - 1$, find $\{n_{h,\varepsilon}^{k+1}, \varphi_{h,\varepsilon}^{k+1}\}$ in $V^h \times V^h$ such that $\forall \chi \in V^h$ we have

$$\begin{cases} \left(\dfrac{n_{h,\varepsilon}^{k+1} - n_h^k}{\Delta t}, \chi\right)^h + \left(\tilde{B}_\varepsilon(n_{h,\varepsilon}^k)\nabla\left(\varphi_{h,\varepsilon}^{k+1} + \pi^h(\psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1}))\right), \nabla\chi\right) = 0\,, & \text{(3.75a)} \\ \sigma\left(\nabla\varphi_{h,\varepsilon}^{k+1}, \nabla\chi\right) + \left(\varphi_{h,\varepsilon}^{k+1}, \chi\right)^h = \gamma\left(\nabla n_{h,\varepsilon}^{k+1}, \nabla\chi\right) + \left(\overline{\psi}'_-(n_h^k - \dfrac{\sigma}{\gamma}\varphi_h^k), \chi\right)^h\,, & \text{(3.75b)} \end{cases}$$

where $\tilde{B}_\varepsilon$ is defined by the above upwind method for the regularized mobility. Indeed, for the definition of the upwind coefficient $B_{\varepsilon,ij}^k$ involved in the calculation of the matrix U_ε associated with $\left(\tilde{B}_\varepsilon(n_{h,\varepsilon}^k)\nabla\cdot, \nabla\cdot\right)$ a slight modification has to be made compared to (3.73). The regularized upwind mobility $B_{\varepsilon,ij}^k$ is given by (3.73) if $n_{\varepsilon,i}^k$ and $n_{\varepsilon,j}^k$ are in $]\varepsilon, 1-\varepsilon[$. For $n_{\varepsilon,i}^k \leq \varepsilon$, it is replaced in (3.73) by ε and if $n_{\varepsilon,i}^k \geq 1-\varepsilon$, it is then replaced by $1-\varepsilon$ (the same applies for $n_{\varepsilon,j}^k$). Altogether, we obtain that there are two positive constants b_1 and B_1, such that for each pair of nodes

$$b_1 < B_{\varepsilon,ij}^k < B_1,$$

and $s \in V^h$, we have

$$b_1 < \tilde{B}_\varepsilon(s) < B_1.$$

The principal difference with the scheme (3.28a)–(3.28b) is that here we compute the approximation of the mobility using the previously described upwind approach and this mobility is computed from the solution at the previous time step.

We now prove that this scheme is well-posed, preserves the non-negativity of the discrete solution, is energy stable and, for the regularized case, the discrete solutions converge to the solution of the continuous regularized problem (3.26).

Well-posedness, non-negativity preserving property. We state the following theorem.

Theorem 23 (Existence of non-negative solution and energy stability) *Let $d \leq 3$, and let $\mathcal{T}^h$ be an quasi-uniform acute mesh of the domain Ω. We write the condition*

$$\frac{(d+1)\,G_h\,\Delta t}{\kappa_h^2} \max_{\substack{x_i \in J \\ x_j \in \Lambda_i}} \left(\left|\xi_j^{k+1} - \xi_i^{k+1}\right|\right) < 1, \tag{3.76}$$

where Λ_i is the set of nodes connected to the node $x_i \in J$ by an edge, $G_h = \max_{x_i \in J} |\Lambda_i|$, and $\xi_i^{k+1} = \left(\pi^h\left(\psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right) + \varphi_{h,\varepsilon}^{k+1}\right)(x_i)$.

If the previous condition is satisfied, the system (3.75a)–(3.75b) with an initial condition satisfying $n_h^0 \in K^h$, has a solution $\{n_{h,\varepsilon}^{k+1}, \varphi_{h,\varepsilon}^{k+1}\} \in V^h \times V^h$ where $0 \leq n_{h,\varepsilon}^{k+1} < 1$.

Furthermore, the solutions satisfy the energy inequality

$$E(n_{h,\varepsilon}^{k+1}, \varphi_{h,\varepsilon}^{k+1}) + \Delta t \sum_{k=0}^{N_T-1} \int_\Omega \tilde{B}_\varepsilon(n_{h,\varepsilon}^k) \left|\nabla\left(\varphi_{h,\varepsilon}^{k+1} + \pi^h\left(\psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right)\right)\right|^2 dx \leqslant E(n_h^0, \varphi_h^0)\,, \tag{3.77}$$

where

$$E(n_{h,\varepsilon}^{k+1},\varphi_{h,\varepsilon}^{k+1}) := \frac{\gamma}{2}\left|n_{h,\varepsilon}^{k+1}-\frac{\sigma}{\gamma}\varphi_{h,\varepsilon}^{k+1}\right|_1^2+\frac{\sigma}{2\gamma}\left\|\varphi_{h,\varepsilon}^{k+1}\right\|_0^2 + \left(\psi_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})+\overline{\psi}_-\left(n_{h,\varepsilon}^{k+1}-\frac{\sigma}{\gamma}\varphi_{h,\varepsilon}^{k+1}\right),1\right)^h. \tag{3.78}$$

Proof. *1. Existence of solutions.* To prove the existence of discrete solutions for the system (3.75a)–(3.75b), it is just necessary to adapt the analysis made in Theorem 18. Therefore, we start by defining $w_h^k = n_h^k - \alpha$, where $\alpha := \frac{1}{|\Omega|}\int_\Omega n_h^0$. Using this notation, we rewrite the system of equations (3.75a)–(3.75b) to obtain

$$\begin{aligned}\left(w_{h,\varepsilon}^{k+1}-w_h^k,\chi\right)^h &= -\Delta t\left(\tilde{B}_\varepsilon(w_{h,\varepsilon}^k+\alpha)\nabla\left(\varphi_{h,\varepsilon}^{k+1}+\pi^h\left(\psi'_{+,\varepsilon}(w_{h,\varepsilon}^{k+1}+\alpha)\right)\right),\nabla\chi\right),\\ \left(\varphi_{h,\varepsilon}^{k+1},\chi\right)^h &= \gamma\left(\nabla\left(w_{h,\varepsilon}^{k+1}-\frac{\sigma}{\gamma}\varphi_{h,\varepsilon}^{k+1}\right),\nabla\chi\right)+\left(\overline{\psi}'_-(w_h^k+\alpha-\frac{\sigma}{\gamma}\varphi_h^k),\chi\right)^h.\end{aligned}$$

Then, we define the application $R:\tilde{K}^h\to\tilde{K}^h$, where

$$\tilde{K}^h = \{w\in V^h | M_l\underline{w}\cdot(1,\dots,1)=0\},$$

and is a convex subspace of V^h. Then, we give the application R that reads for $w\in\tilde{K}^h$

$$\begin{aligned}-R(w)=F(w) =& \left[\left(\Delta t M_l^{-1}U_\varepsilon(\underline{w}+\underline{\alpha})\left(M_l+\sigma Q\right)^{-1}(\gamma Q)\right)\right]\underline{w}\\ &+\Delta t M_l^{-1}U_\varepsilon(\underline{w}+\underline{\alpha})\left(\psi'_{+,\varepsilon}(\underline{w}+\underline{\alpha})\right)\\ &+\Delta t M_l^{-1}U_\varepsilon(\underline{w}+\underline{\alpha})\left(M_l+\sigma Q\right)^{-1}\underline{r}^k-\underline{w}^k,\end{aligned}$$

where $\underline{r}^k$ is the vector associated with $M_l\left(\overline{\psi}'_-(\underline{w}_h^k+\underline{\alpha}-\frac{\sigma}{\gamma}\underline{\varphi}_h^k)\right)$. Then, at this state the proof is exactly the same as in the proof of Theorem 18 since $\tilde{B}(s_h)$ is bounded for all $s_h\in V^h$. Therefore, we refer the reader to the other proof for the details of the calculations to prove that R is a Lipschitz continuous application defined on a convex set, and, hence, applying the Brouwer fixed point theorem, we obtain the existence of a solution $\{n_{h,\varepsilon}^{k+1},\varphi_{h,\varepsilon}^{k+1}\}\in V^h\times V^h$ of (3.75a)–(3.75b).

2. Proof of $0\le n_{h,\varepsilon}^{k+1}<1$. To prove the non-negativity of the discrete solution $n_{h,\varepsilon}^{k+1}$, we must have

$$\underline{n}_\varepsilon^{k+1} = \underline{n}^k - M_l^{-1}\Delta t U\left(\underline{\psi}'_{+,\varepsilon}+\underline{\varphi}_\varepsilon^{k+1}\right)\ge 0.$$

Then, using the fact that U is a zero row sum matrix, the previous condition reads for every node $x_i\in J$

$$n_{\varepsilon,i}^{k+1} = n_i^k - \frac{\Delta t}{|D_i|}\sum_{j\in\Lambda_i}U_{ij}\left(\xi_j^{k+1}-\xi_i^{k+1}\right)\ge 0,$$

where we recall that $\xi_i^{k+1} = \left(\pi^h\left(\psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right)+\varphi_{h,\varepsilon}^{k+1}\right)(x_i)$, $|D_i|$ denotes the volume of the barycentric cell associated to the node x_i, and Λ_i is the set of node connected to the node x_i by an edge. Then, from the fact that the mesh is acute and the mobility is positive, we know that each non-diagonal entry U_{ij} is non-positive. Using the definition of the upwind mobility B_{ij}^k

defined by (3.73), we have

$$U_{ij}\left(\xi_j^{k+1}-\xi_i^{k+1}\right)=Q_{ij}\left(n_i^k(1-n_j^k)^2\min(0,\xi_j-\xi_i)+n_j^k(1-n_i^k)^2\max(0,\xi_j-\xi_i)\right).$$

Therefore, to preserve the non-negativity, we only need to focus on the case $\xi_j-\xi_i<0$, leading to the condition

$$n_i^k-\frac{\Delta t}{|D_i|}G_h\,|Q_{ij}|\max_{\substack{x_i\in J\\ x_j\in\Lambda_i}}\left(\left|\xi_j^{k+1}-\xi_i^{k+1}\right|\right)\geq 0,$$

where G_h is the maximum number of connected nodes, i.e. $G_h=\max_{x_i\in J}|\Lambda_i|$. Then, from [94], we know that

$$\frac{|Q_{ij}|}{M_{l,ii}}\leq\frac{Q_{ii}}{M_{l,ii}}\leq\frac{(d+1)}{\kappa_h^2},$$

where we recall that κ_h is the minimal perpendicular length found for all elements $K\in\mathcal{T}^h$. Using this property, we recover the condition to obtain the non-negativity of the discrete solution

$$\frac{(d+1)\,G_h\,\Delta t}{\kappa_h^2}\max_{\substack{x_i\in J\\ x_j\in\Lambda_i}}\left(\left|\xi_j^{k+1}-\xi_i^{k+1}\right|\right)\leq 1.$$

Then, the upper bound for n_h^{k+1} can be found using the same approach. Indeed, we search to satisfy the condition

$$n_i^k-\frac{\Delta t}{|D_i|}\sum_{j\in\Lambda_i}U_{ij}\left(\xi_j^{k+1}-\xi_i^{k+1}\right)<1,$$

for all node $x_i\in J$. Repeating the same kind of calculations as for the lower bound, we find

$$\frac{\Delta t}{|D_i|}G_h\,|Q_{ij}|\max_{\substack{x_i\in J\\ x_j\in\Lambda_i}}\left(\left|\xi_j^{k+1}-\xi_i^{k+1}\right|\right)<1-n_i^k.$$

Using the definition of the upwind mobility in the case $\xi_j^{k+1}-\xi_i^{k+1}\geq 0$, we obtain the condition

$$\frac{(d+1)\,G_h\,\Delta t}{\kappa_h^2}\max_{\substack{x_i\in J\\ x_j\in\Lambda_i}}\left(\left|\xi_j^{k+1}-\xi_i^{k+1}\right|\right)<1,$$

and we retrieve the strict inequality in the stability condition (3.76).

3. Energy stability. The energy (3.78) and its dissipation (3.77) are found from the same calculation as in the proof of Theorem 18, and we do not repeat them here.

□

Remark 24 *We want to highlight the fact that even if we presented the proof of the existence of solutions, the non-negativity preservation, and the energy stability in the regularized case, these results hold for the non-regularized case.*

Stability and convergence. Since we work in the regularized context for the upwind scheme, the energy estimate (3.78)–(3.77) provide us sufficient inequalities to analyze the convergence of the scheme toward the continuous regularized model. We state the following stability inequality established from the energy estimate.

Proposition 25 (Stability bounds for the implicit upwind scheme) *The solution* $\{n_{h,\varepsilon}^{k+1},\varphi_{h,\varepsilon}^{k+1}\}$ *of problem* (3.75a)–(3.75b) *defined by Theorem 23 satisfies*

$$\begin{aligned}\max_{k=0\to N_T-1}\Big(\Big\|n_{h,\varepsilon}^{k+1}\Big\|_1^2+\Big\|\varphi_{h,\varepsilon}^{k+1}\Big\|_1^2\Big)+(\Delta t)^2\sum_{k=0}^{N_T-1}\left(\left\|\frac{n_{h,\varepsilon}^{k+1}-n_h^k}{\Delta t}\right\|_1^2+\left\|\frac{\varphi_{h,\varepsilon}^{k+1}-\varphi_h^k}{\Delta t}\right\|_1^2\right)\\ +\sum_{k=0}^{N_T-1}\Delta t\left|\left(\tilde{B}_\varepsilon(n_h^k)\right)^{\frac{1}{2}}\nabla\left(\varphi_{h,\varepsilon}^{k+1}+\pi^h\left(\psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right)\right)\right|_0^2\\ +\sum_{k=0}^{N_T-1}\Delta t\left(\tilde{B}_{\varepsilon,max}\right)^{-1}\left|\hat{\mathcal{G}}^h\left[\frac{n_{h,\varepsilon}^{k+1}-n_h^k}{\Delta t}\right]\right|_1^2\leq C(n^0),\end{aligned}\tag{3.79}$$

where $\tilde{B}_{\varepsilon,max}=\max_{s\in[0,1)}\tilde{B}_\varepsilon(s)$.

Proof. The first two terms of the inequality (3.79) are obtained from the energy inequality (3.77) and the fact that the upwind mobility is strictly positive. Indeed, to prove that $\left|\nabla n_{h,\varepsilon}^{k+1}\right|_1$ is bounded, we compute

$$\begin{aligned}\min_{n_{h,\varepsilon}^{k+1}}\left(\frac{1+\frac{\sigma}{\gamma}\psi''_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})}{\psi''_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})}\right)^2\left|\psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right|_1^2&\leq\int_\Omega\left(\frac{1+\frac{\sigma}{\gamma}\psi''_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})}{\psi''_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})}\right)^2\left|\nabla\psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right|^2\\ &\leq\left|n_{h,\varepsilon}^{k+1}+\frac{\sigma}{\gamma}\psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right|_1^2.\end{aligned}$$

Therefore, for using an arbitrarily positive parameter $\theta>0$, we have

$$\begin{aligned}\left(\left(\frac{\sigma}{\gamma}\right)^2+\theta\right)\left|\psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right|_1^2\leq\int_\Omega\Big|\nabla\left(n_{h,\varepsilon}^{k+1}-\frac{\sigma}{\gamma}\varphi_{h,\varepsilon}^{k+1}\right)+\frac{\sigma}{\gamma}\nabla\left(\varphi_{h,\varepsilon}^{k+1}+\pi^h\left(\psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right)\right)\\ +\frac{\sigma}{\gamma}\nabla\left(\psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})-\pi^h\left(\psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right)\right)\Big|^2.\end{aligned}$$

Then, we use the argument that the mobility is positive and, hence, using the energy estimate (3.77), we obtain

$$\begin{aligned}\left(\left(\frac{\sigma}{\gamma}\right)^2+\theta\right)\left|\psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right|_1^2&\leq C+\left(\frac{\sigma}{\gamma}\right)^2\left|\left(\psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})-\pi^h\left(\psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right)\right)\right|_1^2\\ &\leq C+\left(\frac{\sigma}{\gamma}\right)^2\left|\psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right|_1^2,\end{aligned}$$

and we proved that

$$\left|\nabla\psi'_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right|_1^2\leq C,$$

from which we obtain

$$\min\left(\psi''_{+,\varepsilon}(n_{h,\varepsilon}^{k+1})\right)\left|n_{h,\varepsilon}^{k+1}\right|_1^2\leq C.$$

The previous inequality is useful since we know that $\min \psi''_{+,\varepsilon}(n^{k+1}_{h,\varepsilon}) > 0$. Therefore, since we have a bound for $\left|n^{k+1}_{h,\varepsilon}\right|_1^2$ alone and using the energy estimate, we obtain

$$\left|\varphi^{k+1}_{h,\varepsilon}\right|_1^2 \leq C.$$

Then, since the previous two inequalities hold for any $k = 0, \dots, N_T - 1$, we obtain the first two terms in (3.79).

The other terms are obtained in the same way as in the proof of Theorem 21, the only difference being to replace the matrix mobility $\tilde{M}_\varepsilon$ by the upwind mobility $\tilde{B}_\varepsilon$. Since the steps of the calculations are the same, we do not repeat the proof here.

□

We now have everything to state our result of convergence. We borrow the notation from section 3.4.4 to define the time interpolants $U_{h,\varepsilon}, W_{h,\varepsilon}$, and the piecewise constant-in-time functions $U^+_{h,\varepsilon}, U^-_{h,\varepsilon}, W^+_{h,\varepsilon}, W^-_{h,\varepsilon}$.

Theorem 26 (Convergence $h, \Delta t \to 0$, $\varepsilon > 0$) *Let $d = 1, 2, 3$ and $n^0 \in H^1(\Omega)$, with $0 \leq n^0 < 1$ a.e. Ω. We assume that $\{\mathcal{T}^h, n^0_h, \Delta t\}_{h>0}$ satisfy*

i) $n^0_h \in V^h$ with $0 \leq n^0_h < 1$, given by $n^0_h = \pi^h(n^0)$ if $d = 1$, and $\hat{P}_h(n^0)$ if $d = 2, 3$.

ii) Let $\Omega \subset \mathbb{R}^d$ be a polyhedral domain and $\mathcal{T}^h$ a quasi-uniform acute mesh.

Therefore, for $\Delta t, h \to 0$ and $\varepsilon > 0$, it exists a subsequence of solutions $\{U_{h,\varepsilon}, W_{h,\varepsilon}\}$ and a pair of function, such that if $d = 1$

$$\{n_\varepsilon, \varphi_\varepsilon\} \in L^\infty\left(0, T; H^1(\Omega)\right) \cap C^{\frac{1}{2},\frac{1}{8}}_{x,t}(\overline{\Omega}_T) \cap H^1(0, T; (H^1(\Omega))') \times L^\infty\left(0, T; H^1(\Omega)\right),$$

and if $d = 2, 3$

$$\{n_\varepsilon, \varphi_\varepsilon\} \in L^\infty\left(0, T; H^1(\Omega)\right) \cap H^1(0, T; (H^1(\Omega))') \times L^\infty\left(0, T; H^1(\Omega)\right),$$

such that

$$U_{h,\varepsilon}, U^+_{h,\varepsilon}, U^-_{h,\varepsilon} \to n_\varepsilon, \quad \text{uniformly on } \overline{\Omega}_T, \quad \text{if} \quad d = 1, \tag{3.80}$$

$$U_{h,\varepsilon}, U^+_{h,\varepsilon}, U^-_{h,\varepsilon} \to n_\varepsilon, \quad \text{strongly in } L^\infty\left(0, T; L^2(\Omega)\right), \quad \text{if} \quad d = 2, 3, \tag{3.81}$$

$$U_{h,\varepsilon}, U^+_{h,\varepsilon}, U^-_{h,\varepsilon} \rightharpoonup n_\varepsilon, \quad \text{weakly in } L^\infty\left(0, T; H^1(\Omega)\right), \tag{3.82}$$

$$\frac{\partial U_{h,\varepsilon}}{\partial t} \rightharpoonup \frac{\partial n_\varepsilon}{\partial t}, \quad \text{weakly in } L^2\left(0, T; \left(H^1(\Omega)\right)'\right), \tag{3.83}$$

$$W_{h,\varepsilon}, W^+_{h,\varepsilon}, W^-_{h,\varepsilon} \rightharpoonup \varphi_\varepsilon, \quad \text{weakly in } L^\infty\left(0, T; H^1(\Omega)\right). \tag{3.84}$$

Moreover, for $d = 1$, $\{n_\varepsilon, \varphi_\varepsilon\}$ is a solution of the regularized-relaxed degenerate Cahn-Hilliard model under the weak form

$$\begin{cases} \int_0^T < \chi, \partial_t n_\varepsilon > &= -\int_{\Omega_T} b_\varepsilon(n_\varepsilon) \nabla\left(\varphi_\varepsilon + \psi'_{+,\varepsilon}(n_\varepsilon)\right) \cdot \nabla\chi, \quad \forall \chi \in L^2(0, T; H^1(\Omega)), \\ \int_{\Omega_T} \varphi_\varepsilon \chi &= \int_{\Omega_T} \gamma \nabla\left(n_\varepsilon - \frac{\sigma}{\gamma}\varphi_\varepsilon\right) \cdot \nabla\chi + \psi'_-\left(n_\varepsilon - \frac{\sigma}{\gamma}\varphi_\varepsilon\right)\chi, \quad \forall \chi \in L^2(0, T; H^1(\Omega)). \end{cases} \tag{3.85}$$

Proof. The proof of the weak and strong convergences (3.80)–(3.84) follows from the same arguments as in Theorem 22 and using the stability inequality (3.79).

Then, to obtain the limit system, the difference with the proof of Theorem 22 is the convergence of the approximation of the continuous mobility. Indeed, since, $\tilde{B}$ is a piecewise constant approximation of b on all $\tilde{D}_{ij}$ on all elements of the mesh, we know that $\tilde{B}(\cdot) \to b(\cdot)$ uniformly and we obtain

$$\int_\Omega \left(\tilde{B}_\varepsilon(U_h^+) - b_\varepsilon(U_h^+)\right) \nabla \left(W_{h,\varepsilon}^+ + \pi^h\left(\psi'_{+,\varepsilon}(U_{h,\varepsilon}^+)\right)\right) \nabla \pi^h \eta \, \mathrm{d}x \to 0, \quad \text{as} \quad h \to 0, \tag{3.86}$$

using a generalized version of the Lebesgue dominated convergence theorem. Therefore, combining (3.74), the previous convergence (3.86), the strong convergence (3.81), the weak convergence (3.84), and (3.15), we obtain

$$\int_\Omega \tilde{B}(U_h^+) \nabla W_h^+ \nabla \pi^h \eta \, \mathrm{d}x \to \int_\Omega b(n) \nabla \varphi \nabla \eta \, \mathrm{d}x.$$

Then, from combining again (3.74), the convergence (3.86), the strong convergence (3.81), and (3.15), we obtain

$$\int_\Omega \tilde{B}(U_h^+) \nabla \pi^h \psi'_+(U_h^+) \nabla \pi^h \eta \, \mathrm{d}x \to \int_\Omega b(n) \psi''_+(n) \nabla n \nabla \eta \, \mathrm{d}x.$$

From the previous results, we show that if we take $\chi = \pi^h \eta$ with $\eta \in H^1\left(0, T; H^1(\Omega)\right)$ in (3.49a), the right-hand side converges to the right-hand side of the first equation of the expected limit system.

The rest of the proof is the same as the proof of Theorem 22, and we do not repeat the calculations here.

□

Remark 27 *Our result of convergence for the upwind scheme is restricted to the regularized case since we do not have an entropy estimate for this definition of the discrete mobility. However, this result is not pessimistic since we know that the regularized model converges to the non-regularized problem as $\varepsilon \to 0$. Furthermore, we want to stress that the non-negativity property and the energy stability is also retrieved for the same scheme applied to the non-regularized version of the model.*

3.6 Linearized semi-implicit numerical scheme

To restrain the computational time of the simulation of the RDCH model within reasonable bounds, we propose a linearized semi-implicit version of our numerical scheme. We linearize the problem using a particular time discretization. The problem now reads:

For each $k = 0, \ldots, N_T - 1$, find $\{n_h^{k+1}, \varphi_h^{k+1}\}$ in $K^h \times V^h$ such that

$$\left(\frac{n_h^{k+1} - n_h^k}{\Delta t}, \chi\right)^h + \left(b(n_h^k)\psi''_+(n_h^k)\nabla n_h^{k+1}, \nabla \chi\right) = -\left(\tilde{B}(n_h^k)\nabla \varphi_h^{k+1}, \nabla \chi\right), \quad \forall \chi \in V^h, \tag{3.87a}$$

$$\sigma\left(\nabla \varphi_h^{k+1}, \nabla \chi\right) + \left(\varphi_h^{k+1}, \chi\right)^h = \gamma\left(\nabla n_h^k, \nabla \chi\right) + \left(\psi'_-(n_h^k - \frac{\sigma}{\gamma}\varphi_h^k), \chi\right)^h, \quad \forall \chi \in V^h. \tag{3.87b}$$

We define the following finite elements matrices

$$U_{ij} = \int_\Omega B_{ij}^k \nabla \chi_i \nabla \chi_j \, \mathrm{d}x, \quad \text{for } i, j = 1, \ldots, N_h, \tag{3.88}$$

and

$$D_{ij} = \int_\Omega b(n_{h,\epsilon}^k)\psi_+''(n_h^k)\nabla\chi_i\nabla\chi_j \,\mathrm{d}x, \quad \text{for } i,j = 1,\dots,N_h. \tag{3.89}$$

We write the matrix form of the equation (3.87a)

$$(M_l + \Delta t D)\,\underline{n}^{k+1} = -\Delta t U \underline{\varphi}^{k+1} + M_l \underline{n}^k,$$

and since U has zero row sum, we can rewrite the previous equation for each node i

$$M_{l,ii} n_i^{k+1} = M_{l,ii} n_i^k - \Delta t \sum_{x_j\in\Lambda_i} \left[D_{ij}(n_j^{k+1} - n_i^{k+1}) + U_{ij}(\varphi_j^{k+1} - \varphi_i^{k+1})\right], \tag{3.90}$$

where Λ_i is the set of nodes connected to the node i by an edge. In the definition of (3.88) we compute the mobility coefficient in function of the direction of $\nabla\varphi_h^k$. As for the nonlinear case, the mobility coefficient is given by

$$B_{ij}^k = \begin{cases} n_i^k(1-n_j^k)^2, & \text{if} \quad \varphi_i^{k+1} - \varphi_j^{k+1} > 0, \\ n_j^k(1-n_i^k)^2, & \text{otherwise.} \end{cases}$$

Even though we cannot redo the same analysis as for the nonlinear scheme and derive the discrete energy, we can establish the existence and the nonnegativity of discrete solutions of (3.87a) and (3.87b).

Theorem 28 (Well-posedness of linear upwind scheme) *Let $\Omega \subset \mathbb{R}^d$, $d = 1,2,3$, and assume that $\mathcal{T}^h$ is a quasi-uniform acute mesh of Ω, and the condition*

$$\frac{(d+1)\,G_h\,\Delta t}{\kappa_h^2} \max_{\substack{x_i\in J \\ x_j\in\Lambda_i}} \left(\varphi_j^k - \varphi_i^k\right) < 1, \tag{3.91}$$

(where Λ_i is the set of node connected to the node x_i by an edge) is satisfied. Then, the linear finite element scheme (3.87a)–(3.87b) with initial condition $n_h^0 \in K^h$ admits a solution $\{n_h^{k+1}, \varphi_h^{k+1}\} \in K^h \times V^h$ satisfying

$$0 \leqslant n_h^{k+1} < 1.$$

Proof. *Step 1. Existence of a unique solution.* Assuming that $\{n_h^k, \varphi_h^k\} \in K^h \times V^h$, from the Lax-Milgram theorem, it exists a unique solution $\varphi_h^{k+1} \in V^h$ of (3.87b) and equation (3.87a) admits a unique solution $n_h^{k+1} \in V^h$. Therefore, it exists a unique pair of discrete solutions $\{n_h^{k+1}, \varphi_h^{k+1}\} \in V^h \times V^h$ for the system (3.87a)–(3.87b). Next, we need to prove that n_h^{k+1} is nonnegative and bounded from above by 1.

Step 2. Nonnegativity and upper bound on n_h^{k+1} for $d = 1,2,3$. First, from the fact that $(M_l + \Delta t D)$ is a M-matrix, we know that its inverse is non-negative, i.e. $(M_l + \Delta t D)^{-1} \geq 0$. Therefore, to preserve the non-negativity of n_h^{k+1}, we need that

$$M_l \underline{n}^k - \Delta t U \underline{\varphi}^{k+1} \geq 0.$$

For every node x_i in $\mathcal{T}^h$, the previous condition reads

$$|D_i|\, n_i^k - \Delta t \sum_{j\in\Lambda_i} B_{ij}^k Q_{ij} \left(\varphi_i^{k+1} - \varphi_j^{k+1}\right) \geq 0,$$

where Λ_i is the set of node connected to the node x_i by an edge. From the fact that the mesh is acute, we know that Q_{ij} is negative. Therefore, using the definition of the mobility coefficient (3.73), we need to focus on the case $\varphi_j^{k+1} - \varphi_i^{k+1} < 0$. In that situation, we have

$$n_i^k - \frac{\Delta t}{|D_i|} \sum_{j \in \Lambda_i} n_i^k \left(1 - n_j^k\right)^2 Q_{ij} \left(\varphi_j^k - \varphi_i^k\right) \geq 0.$$

However, we know that [94],

$$\frac{|Q_{ij}|}{M_{l,ii}} \leq \frac{Q_{ii}}{M_{l,ii}} \leq \frac{(d+1)}{\kappa_h^2}.$$

Hence, we find the following condition to ensure the non-negativity of u_h^{k+1}

$$\frac{(d+1)\, G_h\, \Delta t}{\kappa_h^2} \max_{\substack{x_i \in J \\ x_j \in \Lambda_i}} \left(\varphi_j^k - \varphi_i^k\right) \leq 1.$$

Then, we need to prove that for every node $x_i \in J$ we have n_i^{k+1}, we use Varah's bound [197], and write

$$\left\| \left(\frac{M_l}{\Delta t} + D \right)^{-1} \right\|_\infty \leq \frac{\Delta t}{M_{l,ii}}.$$

Therefore, to retrieve the upper bound on the discrete solution, the condition

$$n_i^k - \frac{\Delta t}{|D_i|} \sum_{j \in \Lambda_i} n_j^k \left(1 - n_i^k\right)^2 Q_{ij} \left(\varphi_j^k - \varphi_i^k\right) < 1,$$

as to be satisfied. Note in the previous equation that we have considered the case $\varphi_j^k - \varphi_i^k > 0$ since in the other case the bound will be satisfied trivially. Then, subtracting n_i^k to both sides of the previous inequality, we obtain

$$-\frac{\Delta t}{|D_i|} \sum_{j \in \Lambda_i} n_j^k \left(1 - n_i^k\right) Q_{ij} \left(\varphi_j^k - \varphi_i^k\right) < 1,$$

and we retrieve the same condition than before with a strict inequality.

Altogether, we proved the existence a unique solution $\{n_h^{k+1}, \varphi_h^k\} \in K^h \times S^h$ for the system (3.87a)–(3.87b) with $0 \leq n_h^{k+1} < 1$ if the stability condition (3.91) is satisfied.

3.7 Numerical simulations

Even though we are presenting numerical results obtained using the linear scheme (3.87a)–(3.87b), the evolution of the energy during the simulations is given from the computation of the discrete formulation of the continuous energy

$$E(n_h^{k+1}, \varphi_h^{k+1}) := \int_\Omega \frac{\gamma}{2} \left| \nabla \left(n_h^{k+1} - \frac{\sigma}{\gamma} \varphi_h^{k+1} \right) \right|^2 + \frac{\sigma}{2\gamma} |\varphi_h^{k+1}|^2 + \psi_+(n_h^{k+1}) + \psi_- \left(n_h^{k+1} - \frac{\sigma}{\gamma} \varphi_h^{k+1} \right) \mathrm{d}x.$$

Table 3.1 – Parameters of the 1D test case

	Parameters
γ	$(0.014)^2$
Δt	0.1γ
δx	0.01
n^0	$\{0.05, 0.3, 0.36\}$
$n^\star$	0.6
σ	5.10^{-5}

First of all, we present test cases in one and two dimensions to validate our method. The physical properties of the solutions such as the shape of the aggregates, the energy decay, the mass preservation and the non-negativity of the solution are the key characteristics we need to observe to validate our method. A comparison with previous results from the literature is also of main importance. The reference used for this study is the work of Agosti *et al.* [8]. The analysis of the long-time behavior of the solutions of the RDCH equation [165] gives us some insights about what we should observe at the end of the simulations. The solutions should evolve to steady-states that are minimizers of the energy functional. Depending on the initial mass, three regions of the cell density should appear. The first being the region of absence of cells, the second the continuous interface linking the bottom and the third, the top of the aggregates. If the initial mass is large enough, the third region is a plateau of the cell density close to the value $n = n^\star$. The study of the effect of the regularization on the numerical scheme is the purpose of the last subsection.

3.7.1 Numerical results: test cases

1D test cases

The table 3.1 summarizes the parameters used for the one dimensional test cases. The initial cell density is a uniformly distributed random perturbation around the value n^0. Figures 3.2 show the evolution in time of the solutions n_h for the three different initial masses.

We can observe that the solution for each of the three test cases remains nonnegative and the mass is conserved throughout the simulations. From figure 3.3, we observe that the energies decrease monotonically for the three simulations but at different speeds. They all display at the end of the simulation a stable (or metastable) state that is a global (or respectively a local) minimizer of the discrete energy.

For the initial condition $n^0 = 0.3$ (Figures 3.2 b1), b2), b3) and Figure 3.3 in the middle), the energy decreases rapidly and reaches a plateau showing that the solution evolves rapidly to a steady state. The solution at $t = 10$ is organized in aggregates that are not saturated (i.e. the maximum density is below $n^\star$). The explanation behind this observation is that the initial mass is not sufficient for the system to produce saturated aggregates. However, the clusters appear to be of similar thickness and are relatively symmetrically distributed in the domain.

For $n^0 = 0.36 > n^\star/2$ (Figures 3.2 c1), c2), c3)), aggregates are thicker. The top of the aggregate located at the center of the domain is flat and reach the maximal value $n^\star$ (which is not the case for the other test cases). Likewise, the symmetry in the domain is respected. Using Figure 3.3 on the right, we observe that at different times, the energy evolves through several meta-stable equilibria. This reflects the fact that the solution reached different meta-stable states before a stable equilibrium that minimizes more the energy.

For the initial condition $n^0 = 0.05$ (Figures 3.2 a1), a2), a3)), the shape of the final solution

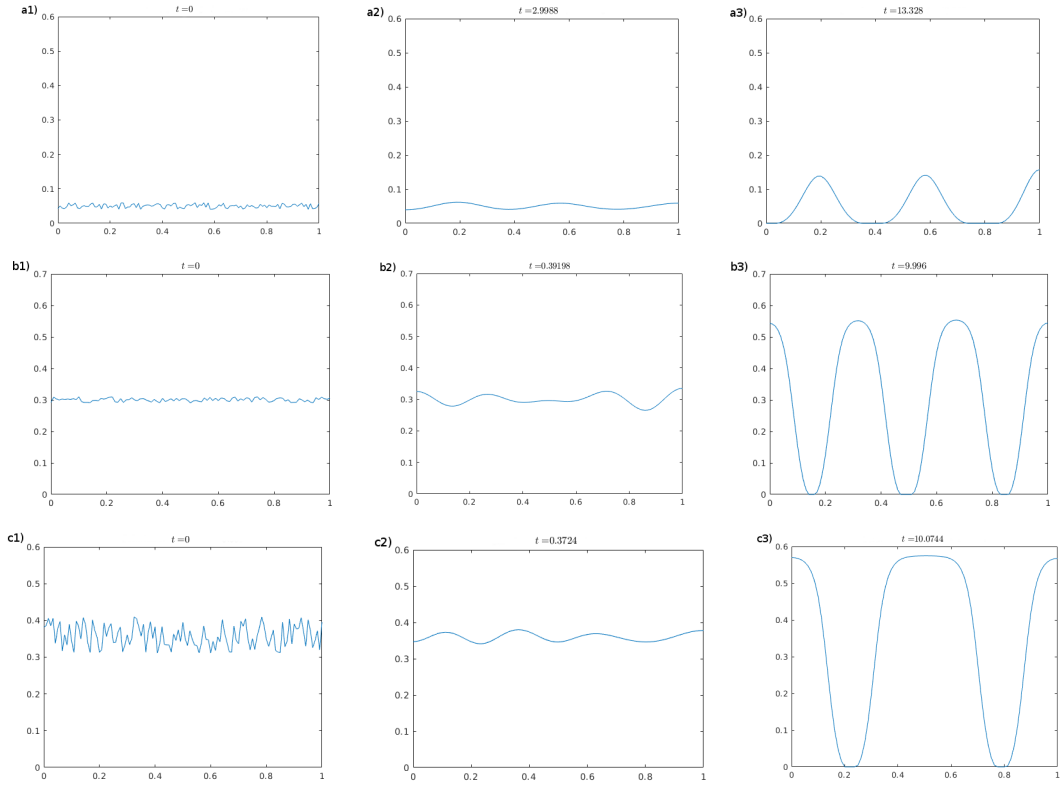


Figure 3.2 – Solution n_h at 3 different times with $n^0 = 0.05$(a1,a2,a3), $n^0 = 0.3$ (b1,b2,b3) and $n^0 = 0.36$ (c1,c2,c3).

is different. The aggregates appear to be thinner and far from each other. The symmetry is not retrieved in the domain. Furthermore, from Figure 3.3 (on the left), we can observe that the evolution of the solution is slow compared to the two other test cases. The evolution of the energy in the first moments of the simulation is slow. This first moments correspond to the spinodal decomposition phase. The slow evolution of the solution is explained from the fact that the mobility is degenerate and the amount of mass available in the domain is small. Using Figure 3.3, we can also see that the energy continues to decrease even at the end of the simulation. To keep comparable simulation times, we did not reach the complete steady state.

Let us compare qualitatively these results with the ones obtained in [8]. For the two test cases $n^0 = 0.3$ and $n^0 = 0.36$, there is no differences in the shape of the aggregates or in the distribution of the mass in the domain. For $n^0 = 0.05$, some small discrepancies with the final solutions are observed. In particular, the symmetry of the aggregates in the domain is not respected in our case whereas it is in the reference work. We must stress that doing other simulations, the symmetry was sometimes reached at the time $t \approx 100$ for the initial condition $n^0 = 0.05$. The reason is that the system will evolve to respect the symmetry but the time at which this stable-steady state is reached depends on the initial distribution of the cell density.

Altogether, the solutions obtained at the end the three simulations are in accordance with the description of the steady-states made in [165]. The three regions of interest are indeed retrieved at the end on each simulation.

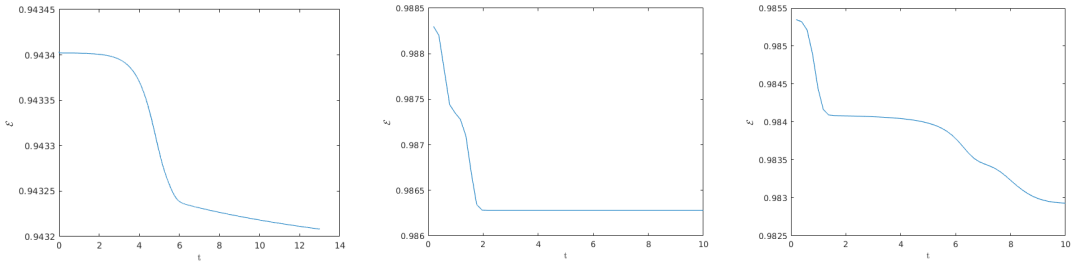


Figure 3.3 – Evolution of discrete energy through time for the 3 initial conditions (from left to right $n^0 = \{0.05, 0.3, 0.36\}$)

Table 3.2 – Parameters of the test cases

	Parameters
γ	0.014^2
Δt	2γ
Δx	$1/64$
n^0	$[0.05, 0.3, 0.36]$
$n^\star$	0.6
σ	10^{-5}

2D test cases

For the two-dimensional test cases, the domain is a square of length $L = 1$. The initial density is computed in the same way as for the one-dimensional test cases, i.e. a random uniformly distributed perturbation around n^0. The summary of the parameters can be found in table 3.2. Figures 3.4, depict the results of three test cases with different initial masses. The three simulations satisfy the nonnegativity of the cell density, the conservation of the initial mass and the monotonic decay of the discrete energy. However, different shapes can be observed for the aggregates.

Figures 3.4 a1),a2),a3) show the evolution of the solution through time for the small initial mass $n^0 = 0.05$. Starting from a uniform random distribution of the cell density in the domain, the solution evolves into a more organized configuration. Progressively, a separation of the two phases of the mixture occurs. At the end of the simulation, small clusters are formed. They display a circular shape and are of similar width. The organization of the clusters in the domain tries to maximize the distance between each others. Using the Figure 3.5 (left), we observe a drop of the energy in the first moments of the simulation denoting a fast reorganization of the random distributed initial condition. Then, the solution appears to evolve very slowly, i.e. a meta-stable state was reached. A second drop of the energy follows around $t \approx 15$, the system enters the "coarsening" phase: the small aggregates become more dense and merge with others. At the end, the evolution is very slow. The system continues to rearrange but due to the degeneracy of the mobility and the small amount of initial mass this process is very slow.

Figures 3.4 b1),b2),b3) show the evolution of the solution for $n^0 = 0.3$. The two successive processes that are the spinodal decomposition and the coarsening are observed. Between the Figures 3.4 b1) and 3.4 b2), we observe that the solution evolves from a random uniform configuration to an organization in small aggregates that are not saturated. Then (Figure 3.4 b3)), the cell density is distributed in elongated and saturated aggregates. The separation of the two

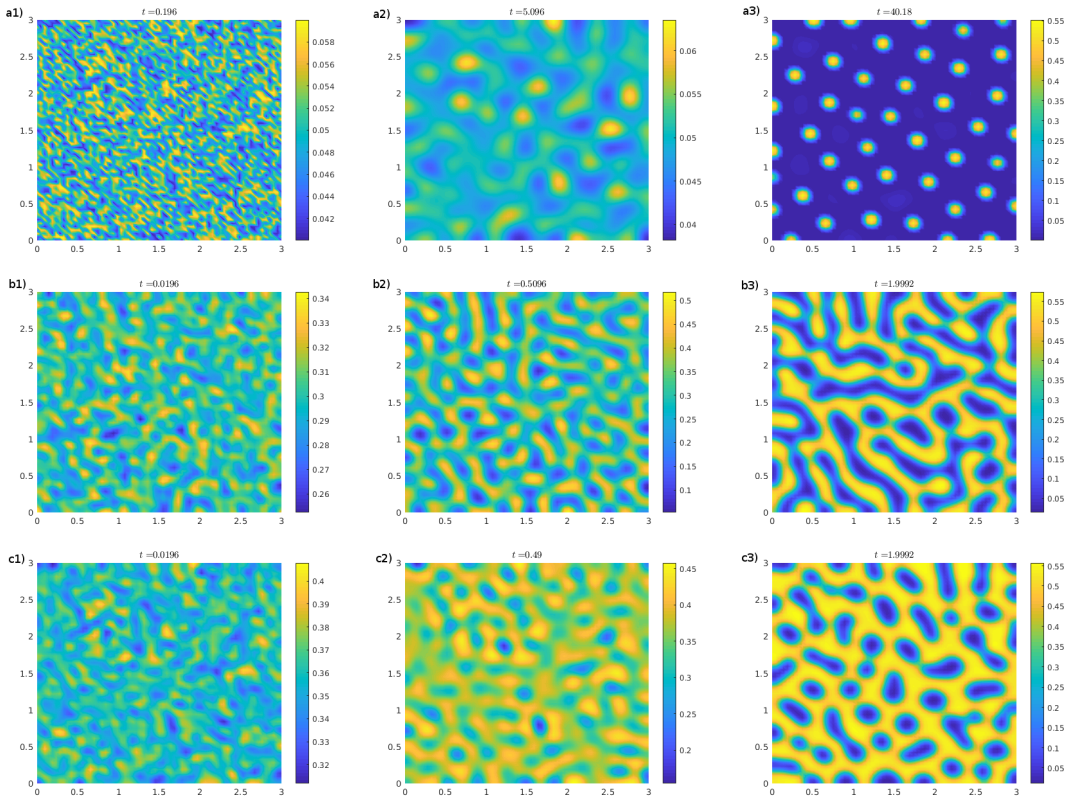


Figure 3.4 – Solution n_h at 3 different times with $n^0 = 0.05$ (a1,a2,a3), $n^0 = 0.3$ (b1,b2,b3) and $n^0 = 0.36$ (c1,c2,c3).

phases is clear. However, using Figure 3.5 (middle), we observe that at the end of the simulation the cell density continues to rearrange. Due to the degeneracy of the mobility, this evolution is again very slow.

On Figures 3.4 c1),c2),c3), we can observe the evolution of the solution for $n^0 = 0.36$. Again, the solution goes through the spinodal decomposition and coarsening phases. The only difference that needs to be highlighted for this simulation is the shape of the aggregates at the end. Indeed, the initial mass being $n^0 = 0.36 > n^\star/2$, aggregates are wider and more connected to each others.

Therefore, depending on the initial mass of cells in the domain, the 2D simulations of the model show very different spatial organizations of the cell density.

Compared to the reference work [8], the organizations of the cells for the different initial cell densities are the same. No clear difference can be established regarding the simulation involving the relaxed model and the original one.

The three regions corresponding to a steady-state described in [165] are retrieved at the end of the simulations for these 2D test cases.

3.7.2 Effect of the relaxation parameter σ

In this section we evaluate the effect of the relaxation parameter σ for the stability of the scheme, and in particular to satisfy the CFL-like condition (3.91). This conditions is necessary

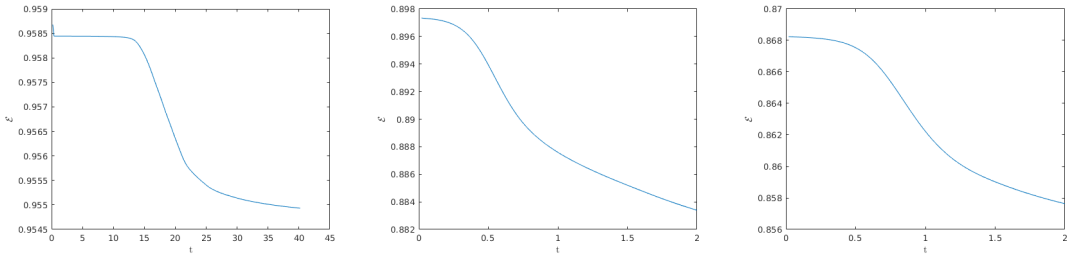


Figure 3.5 – Evolution of discrete energy through time for the 3 initial conditions (from left to right $n^0 = \{0.05, 0.3, 0.36\}$)

to preserve the nonnegativity of the solutions of the linear discrete scheme. To evaluate the effect of this parameter on the choice of the time step Δt, we compute the amplification matrix H defined by

$$X^{k+1} = HX^k, \quad \text{with} \quad X^k = \begin{bmatrix} \underline{n}^k \\ \underline{\varphi}^k \end{bmatrix}.$$

Here, X^k is called the state vector. Using the matrix form of the scheme (3.30)–(3.31) we can decomposed the amplification matrix by $H = H_1^{-1} H_2$ with

$$H_1 = \begin{bmatrix} 0 & \sigma A + M \\ M + \Delta t D & \Delta t U \end{bmatrix}, \qquad H_2 = \begin{bmatrix} \gamma A - (1 - n^\star)M & \frac{\sigma}{\gamma}(1 - n^\star)M \\ M & 0 \end{bmatrix}.$$

We denote by $\lambda_i, i = 1, \ldots, N$, the eigenvalues of the amplification matrix H.

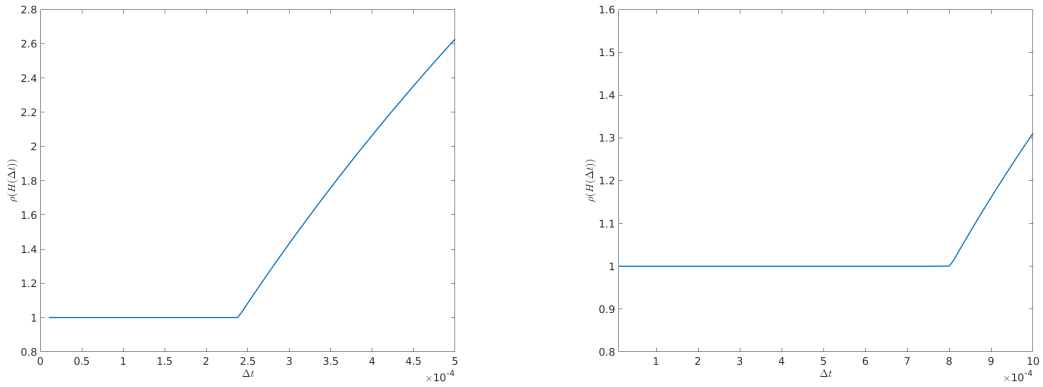
To analyze the stability of the numerical scheme due to the relaxation parameter, we compute the spectral radius of the amplification matrix

$$\rho(H(\Delta t)) = \max_i(|\lambda_i|),$$

for a smooth initial conditions. The scheme is stable when the maximum value of the modulus of the eigenvalues is less or equal to 1. The figure 3.6 represents the spectral radius in function of the time step Δt for two values of σ (the other parameters are the ones taken from the one dimensional test cases with $n^0 = 0.3$). We can observe that the scheme remains stable when Δt is small for the two test cases. However, we see that increasing σ allows to take larger time steps while remaining stable. This result can be explained due to the fact that increasing σ diminishes the value $\max_{\substack{x_i \in J \\ x_j \in \Lambda_i}} \left(\varphi_j^k - \varphi_i^k\right)$ present in the stability condition (3.91). Therefore, the regularization induced by the relaxation parameter allows for faster simulations, but it has an effect on the accuracy of the solution compared to the solution given by the non-relaxed model. However, at the moment it remains unclear how to compare the solution given by a simulation of the relaxed model and a solution of the original degenerate model (without relaxation). This will be the subject of a further work.

3.8 Conclusion

We described and studied two finite element schemes to solve the relaxed degenerate Cahn-Hilliard equation with single-well logarithmic potential. The difference between the two is in the approximation of the continuous mobility function. The first scheme uses the idea proposed in



(a) $\sigma = 10^{-5}$ (b) $\sigma = 10^{-4}$

Figure 3.6 – Spectral radius in function of Δt for $\sigma = 10^{-5}$(left) and $\sigma = 10^{-4}$ (right).

[108] and allows to obtain an entropy estimate from the discrete scheme. The second aims to recover an efficient and practical discretization of the relaxed degenerate Cahn-Hilliard model. The mobility in this scheme is computed from an adaptation of the upwind method in the finite element framework. Even though we cannot prove that the solutions of the upwind scheme convergence to the solutions of the continuous relaxed degenerate Cahn-Hilliard model, we showed that it preserves the non-negativity of the order parameter and the dissipation of the energy. For the scheme that uses the upwind method, we considered two different time discretizations leading to a nonlinear semi-implicit scheme and a linear semi-implicit one.

The linear semi-implicit scheme allows for faster simulations and we proved that it is well-posed and preserves the nonnegativity of the solutions as well. We presented some numerical simulations using this linear scheme in one and two dimensions. The numerical simulations validated the nonnegativity-preserving and energy decaying properties of the scheme. The numerical solutions of the finite element approximation of the RDCH model are in good agreement with previous works dealing with the non-relaxed model. We showed that the relaxation parameter σ allows us to take a larger time step in the scheme (as long as the condition for the nonnegativity is preserved and $\sigma < \gamma$). We point out that thanks to the spatial relaxation, our numerical scheme can be easily implemented and simulations of the relaxed degenerate Cahn-Hilliard model can be computed efficiently using standard softwares.
In a work in preparation, we will study the phase-ordering dynamics of the system and the error analysis of the discrete solutions. An emphasis will be put on the effect of the regularization parameter because, so far, we lack a clear explanation on how it affects the accuracy of solutions compared to the original system.

3.A Proof of M-matrix properties in the 1D and 2D cases

Proposition 29 *For $d = 1, 2$, the matrix $(\frac{M_l}{\Delta t} + D)$ is a M-matrix.*

Proof. If the mass matrix is lumped, the all matrix is a Z-matrix due to the fact that the non-diagonal terms of $L_{b\psi''_+}$ are negative. Therefore, the sum of the lumped mass matrix M_l and

D is a Z-matrix. Furthermore, we can write

$$\frac{M_l}{\Delta t} + D = cI - B,$$

where I is the identity matrix, c is a constant and B is a matrix with $b_{ij} \geq 0, \quad 1 \leq i, j \leq N$. Let us choose $c = \max(\frac{M_{l,ii}}{\Delta t} + (D)_{ii})$ and consequently the matrix B can be deduced and contains only positive terms. Therefore, we have proved that $(\frac{M_l}{\Delta t} + D)$ is a M-matrix.

Part II

Modification of existing nonlinear PDE models, numerical simulation, and application in Biology.

Chapter 4

Treatment-induced shrinking of tumour aggregates: A nonlinear volume-filling chemotactic approach

Abstract

Motivated by experimental observations in 3D/organoid cultures derived from glioblastoma, we develop a mathematical model where tumour aggregate formation is obtained as the result of nutrient-limited cell proliferation coupled with chemotaxis-based cell movement. The introduction of a chemotherapeutic treatment induces mechanical changes at the cell level, with cells undergoing a transition from rigid bodies to semi-elastic entities. We analyse the influence of these individual mechanical changes on the properties of the aggregates obtained at the population level by introducing a nonlinear volume-filling chemotactic system of partial differential equations. The elastic properties of the cells are taken into account through the so-called *squeezing probability*, which allows us to change the packing capacity of the aggregates, depending on the concentration of the treatment in the extracellular microenvironment. We explore two scenarios for the effect of the treatment: firstly, the treatment acts only on the mechanical properties of the cells and, secondly, we assume it also prevents cell proliferation. A linear stability analysis enables us to study the ability of the system to create patterns. We provide numerical simulations in 1D and 2D that illustrate the shrinking of the aggregates due to the presence of the treatment.

This chapter is taken from Luís Almeida, Gissell Estrada-Rodriguez, Lisa Oliver, Diane Peurichard, A. P., Francois Vallette, *Treatment-induced shrinking of tumour aggregates: A nonlinear volume-filling chemotactic approach*, Journal of Mathematical Biology, 83, 29 (2021). Journal.

4.1 Introduction

Cell migration in the extracellular microenvironment (ECM) and the organisation of cells in response to chemical and mechanical cues are successfully studied using continuum descriptions based on differential equations [6, 163]. In a continuous setting, the chemotactic behaviour of cells (*i.e.* the ability to move along a chemical gradient) is often modelled using a Keller-Segel system of equations [122]. This model was originally proposed for pattern formation in bacterial populations but turned out to be pertinent to describe a wide variety of self-organisation behaviours [54, 30, 206, 125, 202]. Different variations of the Keller-Segel model have been adopted in order to better understand the way cells aggregate [44, 14, 160, 184, 69].

Cancer cells have been shown to respond to chemical and mechanical signals from components of the tumour micro-environment (TME) and *vice versa*, cells in the TME acquire a more tumourigenic phenotype, which would sustain tumour growth. The interaction of tumour cells with the TME has been the subject of recent biological surveys [113, 88]. Many *in vitro* (and *ex vivo*) experiments have shown that cells that are cultured on ECM often have a tendency to form aggregate patterns that are dependent on the particular cell lines and physical properties of the media [91]. The exact consequences of the dynamic interplay between heterogeneous cellular entities and their response to alterations in the TME have not yet been elucidated. In particular, little is known about the role of mechanics in the spatial organization of the tumour spheroids. Biological evidence presented in [144, 128, 57] suggests that the formation of aggregates in glioma cells can be explained through a chemotaxis process, rather than, *e.g.*, cell-cell adhesion. In this paper, we follow the chemotactic approach to explain the formation of glioma aggregates. For the case of breast cancer cells, a recent report [45] has proposed a chemotaxis-based explanation for spheroid formation based on theoretical analysis and numerical simulations of the Turing instabilities of such systems.

Inspired by experimental observations in 3D/organoid cultures derived from freshly operated Glioblastoma (GBM), which reproduce *in vivo* behaviours as described in [155] (see Section 4.2 for more details), in this paper we explore a simple setting where GBM aggregate formation is the result of nutrient-limited cell proliferation coupled with a chemotaxis-based cell movement. A chemotherapeutic treatment introduced after the formation of the aggregates induces mechanical changes at the cell level. We study the influence of these individual mechanical changes on the characteristics of the aggregates obtained at the population level.

GBM are solid tumours characterised by intra- and inter-tumoural heterogeneity and resistance to conventional treatments that result in a poor prognosis [140]. They are the most common and aggressive primary brain tumour in adults. Standard treatments include surgical resection (when possible), combined with radiotherapy and chemotherapy using the DNA alkylating agent Temozolomide (TMZ) [187]. In fact, the overall survival of treated patients is about 15 months versus 3 months without treatment, with fewer than 5% of patients surviving longer than 5 years [186].

One reason behind this relative therapeutic failure is the poor response of GBM tumours to this chemotherapeutic treatment due to their plasticity. Several studies have looked at the genetic compounds of TMZ-resistant cells focusing on the genes responsible for DNA mismatch repair protein [189], while other studies focused on spatial and temporal variations in signalling pathways, which lead to functional and phenotypic changes in GBM [157]. The communication between the tumour cells and the TME as well as the properties of the ECM have a large impact on tumour evolution and invasion, as shown in recent studies [68, 40, 205]. From a biological and medical perspective, it is difficult to investigate the connections between clinically observable glioma behaviour and the underlying molecular and cellular processes. The challenge is to integrate the theoretical and empirical acquired knowledge to better understand the mechanisms and factors that contribute to GBM resistance to treatment. In this context, mathematical models provide useful tools towards identifying dependencies between different phenomena and how they are affected by the different therapeutic strategies.

Much effort has been dedicated to the modelling of GBM formation and invasion of the surrounding tissue, as well as to improving diagnosis and treatment. The exhaustive review [11] discusses different modelling approaches as well as some of the main mechanisms that are observed in GBM formation and invasion.

In this work, we explore two scenarios: the case where the treatment only acts on the mechanical properties of the cells, and the case where it also prevents cell proliferation [130]. We adopt a macroscopic approach where cells are represented by their macroscopic density and are

supposed to move in the environment via chemotaxis, *i.e.* towards zones of high concentrations of a chemoattractant that is produced by the tumour cells. Moreover, cell proliferation is assumed to depend on the local concentration of nutrients available. Finally, we suppose that when the treatment - represented by its continuous concentration - is introduced, it diffuses in the environment and is naturally consumed by the cells.

Under these hypotheses, which are motivated by the experimental results discussed in Section 4.2, we obtain a nonlinear volume-filling Keller-Segel model for the cell density, coupled with reaction diffusion equations for the chemoattractant and treatment concentrations. Moreover, we provide a linear stability analysis that enables us to study the ability of the system to generate patterns, and we provide numerical simulations in 1D and 2D.

The paper is organised as follows. In Section 4.2 we describe the *in vitro* experiments and the main experimental observations that motivated our model. Section 4.3 is devoted to the description of the model for the first part of the experiments (without the treatment, Section 4.3.1) and the second part, when the treatment is introduced (Section 4.3.2). In Section 4.4, we present the stability analysis for each of these two parts, and Section 4.5 is devoted to numerical simulations. Section 4.5.2 presents the results in 1D including a discussion on the comparison between numerical experiments and theoretical predictions of the stability analysis. Section 4.5.3 shows the 2D simulations. Finally, we discuss the results and give some perspectives in Section 4.6.

4.2 Description of the experiments

To address the question of the response of GBM cells to TMZ treatment, we took advantage of recently developed 3D biosphere experiments, using GBM patient-derived cultures in a simple 3D scaffold composed of alginate and gelatin [156]. GBMG5 cells were cultured at a concentration of 4×10^6 cells/ml for 14 days until the formation of cell aggregates could be observed, corresponding to the first part of the experiments (**P1**). After that, the cultures were treated with 100 μM TMZ for two hours once every week (second part of the experiments, **P2**). Over the 30 days, the proliferation was determined counting 3 representative samples, and the cell number was determined as follows. The biospheres were dissociated by incubation for 3 min in 100 mM Na-Citrate and the cell number and cell viability were determined using the Countess optics and image automated cell counter (Life Technologies). In addition the aggregates were photographed to analyse their morphology, and the diameter of the cell aggregates were measured from pictographs using FIJI. To determine the diameters of these cell structures, more than 200 cell aggregates were measured.

We show in Figure 4.1 (I) the mean length (in μm) of the cell aggregates computed from the microscopic images without TMZ treatment (round markers), and with TMZ weekly administered (squared markers). Figures 4.1 A and B show typical microscopic images of the spheroids at day 24, without and with weekly TMZ treatment respectively. In Figure 4.1 (II), we show the evolution of the cell number as a function of time, where we do not observe big changes in cell number with and without TMZ, once the carrying capacity of the system is reached.

Using clinically relevant concentrations [170] the total number of cells in the biospheres does not seem to be significantly impacted by the TMZ treatment. However, the mean size of the GBMG5 cell aggregates decreases when TMZ is introduced weekly as compared to control cultures (Figure 4.1 (I)), suggesting that in the presence of the treatment, GBMG5 cells tend to self-organize into smaller and more compact cell clusters.

Another observation supporting a tendency for GBM cells to form more compact structures with higher intracelluar adhesion under this type of treatment is the increased expression of

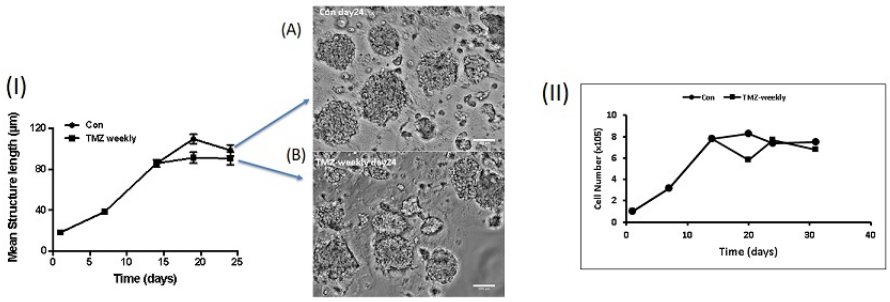
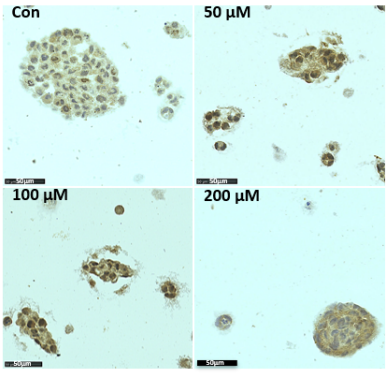


Figure 4.1 – Biological experiments of GBMG5 cells cultured in 3D scaffolds with or without 100μM TMZ. (I) Evolution of the mean cell aggregates diameter determined from the microscopic images as function of time, without treatment (circle markers) and with a weekly TMZ (square markers). (A) Typical microscopic image on day 24 of control cell aggregates without treatment, (B) microscopic image at day 24 with 100μM of TMZ administrated weekly for 2 hours. (II) Evolution of the total cell number in the biospheres as function of time, without treatment (circle markers) and with weekly TMZ (square markers).



	Claudin
control	+
50 µM TMZ	++
100 µM TMZ	+++
200 µM TMZ	+++

Figure 4.2 – Claudin expression marking the tight junctions between the cells.

claudin, a marker of tight junctions formed between cells (Figure 4.2). In Figure 4.2 we have GBMG5 cells that were cultured for 14 days until the formation of cell aggregates could be observed, and the cultures were treated with different doses of TMZ until day 23. Furthermore, cell aggregates were fixed then labeled with an anti-claudin antibody. The degree of staining was determined in a double-blind experiment. As one can observe from the images of Figure 4.2 (left), the cell aggregates seem to be smaller and more compact when TMZ is present compared to the control group (top left figure), and these cultures are associated with higher levels of claudin (see Figure 4.2 (right)).

Inspired by these observations and results presented in [144, 128, 57], we propose a general model for the chemotaxis-driven formation of cell clusters and shrinking of the aggregates via the action of a non-cytotoxic drug whose effect would be characterised by a change in the mechanical properties of cells that would just modify their packing properties. Moreover, we also consider a second scenario where the treatment would also affect cell proliferation and compare the results obtained.

4.3 Mathematical model

Motivated by the experiments described in Section 4.2, we assume that glioma cells have a chemotactic behavior, *i.e.* they move in response to some signaling chemical (chemoattractant), which is secreted by themselves and diffuses in the environment. The chemotactic movement of cell populations plays a fundamental role for example in gastrulation [70], during embryonic development; it directs the movement of immune cells to sites of infection and it is crucial to understand tumour cell invasion [64] and cancer development [206]. Motivated by these applications, chemotaxis and related phenomena have received significant attention in the theoretical community, see the reviews [117, 118].

We suppose that in normal conditions (first part of the experiments described in Section 4.2, where there is no treatment), tumour cells proliferate and move via chemotaxis as described before. We suppose that cell proliferation is limited by the nutrients available in the environment, *i.e.* cell proliferation is only active as long as the local density does not exceed a given threshold corresponding to the carrying capacity of the environment. Moreover, in order to take into account the finite size of cells and volume limitations, cell motion is only allowed in locations where the local cell density is much smaller than another threshold value corresponding to the tight packing state. In normal conditions, cells are supposed to behave as rigid bodies. In stressed conditions however, (second part of the experiments described in Section 4.2) we suppose that cells respond to the chemotherapeutic stress (induced by the presence of the treatment) by changing their mechanical properties and behaving as a semi-elastic material.

These hypotheses are modelled via a system of partial differential equations (PDE) which corresponds to a volume-filling chemotaxis equation [162] for the first part (to describe the self-organization of cells into aggregates), and a "semi-elastic" volume-filling chemotaxis approach [201] for the second part, when the treatment is introduced.

For convenience we denote the density of the population of cancer cells in the first part of the experiments by $u(\mathbf{x}, t)$ and that of the second part by $w(\mathbf{x}, t)$. Here $\mathbf{x} \in \Omega \subset \mathbb{R}^2$ where Ω is a bounded domain. The time $t \in [0, T]$ where $T = T_1 + T_2$ represents the total time corresponding to the first and second parts. The main difference between these two populations is the change in the elastic properties of the cells due to the presence of the treatment. If the concentration of the treatment is zero, $u(\mathbf{x}, t) = w(\mathbf{x}, t)$. Cells follow a biased random walk according to the distribution of the chemoattractant of concentration $c(\mathbf{x}, t)$ that is secreted by the cells. We start by detailing the different components of the mathematical models corresponding to the two parts

of the experiments described in Section 4.2.

Logistic growth model for cell proliferation As previously described, in order to take into account the nutrient-limited population growth, cell proliferation is modelled by a logistic growth process. At the population level, we consider a source term $f(u)$ in the PDE for the evolution of the cell density which depends nonlinearly on the local cell density u and reads:

$$f(u) = ru\Big(1 - \frac{u}{u_{\max}}\Big) \ . \tag{4.1}$$

Here, $r > 0$ is the proliferation rate and $u_{\max}$ corresponds to the maximum density of the population, also referred to as the carrying capacity of the environment. Alternative cell kinetics could be considered. For example, we can assume that the proliferation is also mediated by the chemoattractant concentration such that $f(u, c) = ruc(1 - u/u_{\max})$ [162]. In this paper we focus on the case given by (4.1).

Chemoattractant dynamics As described before, we suppose that cell aggregates spontaneously emerge as the result of a self-organization phenomenon of chemotaxis type. To this aim, we suppose that the cells themselves produce the signaling chemical (chemoattractant) that drives their motion. The chemical secreted is therefore supposed to be continuously produced by the cells at rate $\alpha > 0$ and diffuses in the surrounding environment with diffusion coefficient $d_2 > 0$. It is further assumed that the chemical has a finite lifetime and degrades at rate $\beta > 0$. The evolution of the chemical concentration $c(\mathbf{x}, t)$ is therefore given by the following reaction-diffusion equation

$$\partial_t c = d_2 \Delta c + \alpha u - \beta c \ , \tag{4.2}$$

where u is the cell density.

Treatment dynamics We suppose that the treatment is introduced at the beginning of **P2**. This treatment is supposed to diffuse in the environment with diffusion coefficient d_4, and to be consumed by the cells at rate δ. This is modelled by a reaction-diffusion equation for the drug concentration $M(\mathbf{x}, t)$:

$$\partial_t M = d_4 \Delta M - \delta w \ ,$$

where $w(\mathbf{x}, t)$ represents the cell density corresponding to the second part of the experiments. We consider different initial conditions for the drug: either uniformly distributed in the simulation domain, or introduced in the center as a very steep Gaussian function (see Section 4.5).

4.3.1 Volume-filling approach for chemotaxis: first part P1

The classical Keller-Segel system of equations [122] describes how cells move along the gradient to local maxima of the chemoattractant. At the same time, this chemical, which is produced by the cells, promotes aggregation leading to the so-called "overcrowding scenarios", and eventually, the cell density may blow up in finite time (see the comprehensive reviews [117, 188] and references therein).

In this paper, in order to take into account volume limitations and the finite cell size, we consider a modified version of the Keller-Segel model called the *volume-filling* approach for cell motion. This approach was introduced in [162], where the authors provided a detailed derivation in one dimension as well as a comprehensive numerical study of the model.

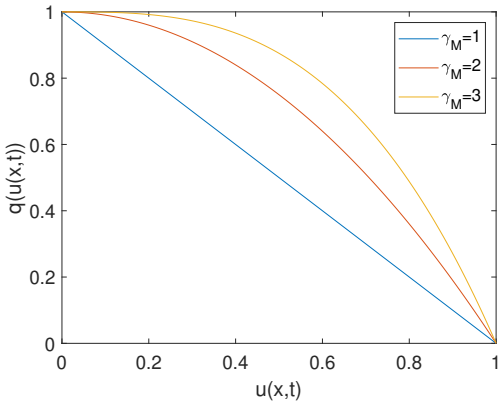


Figure 4.3 – Squeezing probability for different values of γ_M for $\bar{u} = 1$.

Moreover, we introduce the so-called *squeezing probability*, which describes the probability that a cell finds an empty space at a neighbouring location before moving (see [201]). It takes the form

$$q(u(\mathbf{x},t)) = \begin{cases} 1 - \left(\frac{u}{\bar{u}}\right)^{\gamma_M}, & \text{for } 0 \leqslant u \leqslant \bar{u} \ , \\ 0, & \text{otherwise} \ , \end{cases} \tag{4.3}$$

where $u(\mathbf{x},t)$ is the cell density, $\gamma_M \geqslant 1$ is the *squeezing parameter*, $M \equiv M(\mathbf{x},t) \geqslant 0$ denotes the concentration of the treatment, and $\bar{u}$ is the *crowding capacity* which corresponds to the tight packing state of the cells. The function $q(u)$ satisfies the following properties,

$$q(\bar{u}) = 0 \ , \quad 0 < q(u) \leqslant 1 \ , \text{ and } \ q'(u) \leqslant 0 \ \text{ for } \ 0 \leqslant u < \bar{u} \ . \tag{4.4}$$

Moreover, $|q'(u)|$ is bounded and $q''(u) \leqslant 0$.

The exponent γ_M is chosen to be

$$\gamma_M = (\bar{\gamma} - 1)M + 1 \ , \tag{4.5}$$

where $\bar{\gamma}$ is a positive constant, $M = 0$ when there is no drug in the environment (part **P1** of the experiments), and $M \equiv M(\mathbf{x},t)$ when the drug is introduced (part **P2**, described bellow). Such choice of γ_M enables to take into account different forms of squeezing probability, corresponding to different mechanical behaviors of the cells. In Figure 4.3, we plot the squeezing probabilities as a function of the cell density, corresponding to different values of γ_M. We see that when there is no treatment present in the environment ($M = 0$, $\gamma_M = 1$, blue curve of Figure 4.3), the squeezing probability decreases linearly with the local cell density, corresponding to cells modelled as solid particles.

However for larger values of γ_M (when the treatment is present, $M > 0$ and $\gamma_M > 1$, red and yellow curves in Figure 4.3), the squeezing probability becomes a nonlinear function of the cell density, modelling the fact that in the presence of a drug, cells change their mechanical state to behave as semi-elastic entities that can squeeze into empty spaces. We refer to [200] for more details on the link between the cells elastic properties and the squeezing probability.

For convenience of notation, we distinguish the case without treatment ($M = 0$) for **P1** and define

$$q_1(u(\mathbf{x},t)) = 1 - \frac{u}{\bar{u}} \ . \tag{4.6}$$

The Keller-Segel model built with this specific squeezing probability (4.6) has been widely studied in the literature, from modelling [162, 200], analytic [204, 112, 143, 69] and numerical [120] perspectives.

Complete PDE system for the first part P1 Taking into account all the previous ingredients, the complete PDE system for part **P1** (when no treatment is present in the environment) reads:

$$\begin{aligned} \partial_t u &= \nabla \cdot (d_1 D_1(u) \nabla u - \chi_u \phi_1(u) \nabla c) + f(u) \ , \\ \partial_t c &= d_2 \Delta c + \alpha u - \beta c \ , \end{aligned} \tag{4.7}$$

where the first equation describes the evolution of the cell density u and the second is the reaction-diffusion equation for the chemoattractant, previously introduced. The equation for u describes the volume-filling chemotactic motion associated with the squeezing probability q_1. This equation has been obtained as the hydrodynamic limit of the continuous space-time biased random walk model that corresponds to the squeezing probability q_1 (see Appendix 4.A for more details on the derivation). In the hydrodynamic limit, the cell density evolves according to a nonlinear transport diffusion equation with source term, which corresponds to a volume filling Keller-Segel model with logistic growth. As shown in Appendix 4.A, the density-dependent diffusion coefficient $D_1(u)$ and the chemotactic sensitivity $\phi_1(u)$ relate to q_1 via

$$D_1(u) = q_1 - q_1' u \ , \quad \phi_1(u) = q_1 u \ .$$

For this part **P1**, where q_1 is given by (4.6), these coefficients take the values $D_1 = 1$ and $\phi_1(u) = u(1 - u/\bar{u})$. In equation (4.7), $d_1, \chi_u, \alpha, \beta$ are all positive parameters (see Appendix 4.A for the computation of d_1 and χ_u), and $f(u)$ is the logistic growth source term previously mentioned and defined by (4.1).

The PDE system is supplemented with the following zero-flux boundary conditions

$$(d_1 D_1(u) \nabla u - \chi_u \phi_1(u) \nabla c) \cdot \eta = 0 \ , \quad d_2 \nabla c \cdot \eta = 0 \ , \tag{4.8}$$

where η is the outer unit normal at $\partial\Omega$. The initial conditions are given by

$$u(\mathbf{x}, 0) = u_0 \ , \quad c(\mathbf{x}, 0) = c_0 \ . \tag{4.9}$$

4.3.2 PDE system including the treatment: Part P2

We now describe the dynamics of the cell population when the drug is introduced (part **P2** of the experiments described in Section 4.2). As previously mentioned in Section 4.2 and motivated by the observations in [156], where the treatment TMZ does not seem to induce cell death, we suppose that the drug only affects the elastic properties of the cells. For a cell density $w(\mathbf{x}, t)$ the squeezing probability of part **P2** is

$$q_2(w(\mathbf{x}, t), M) = 1 - \left(\frac{w}{\bar{u}}\right)^{\gamma_M} \ . \tag{4.10}$$

Note that we have supposed that the crowding capacity $\bar{u}$ which corresponds to the tight packing state remains unchanged from **P1** to **P2**. This corresponds to the hypothesis that the treatment does not modify the volume of the cells but only change their elastic properties.

The complete PDE system corresponding to the second part of the experiments therefore

reads:

$$\begin{aligned} \partial_t w &= \nabla \cdot (d_3 D_2(w, M)\nabla w - \chi_w \phi_2(w, M)\nabla c) + f(w) \ , \\ \partial_t c &= d_2 \Delta c + \alpha w - \beta c \ , \\ \partial_t M &= d_4 \Delta M - \delta w \ , \end{aligned} \tag{4.11}$$

where $f(w)$ is again the logistic growth source term given by (4.1), and the first equation has been derived using the squeezing probability q_2 defined by (4.10). Note that the carrying capacity $u_{\max}$ remains unchanged in the two parts of the experiments: we have supposed here that the drug does not interact with the nutrients. Analogous to (4.7), the movement of the cells is described by a chemotactic system where, in this case, the diffusion and chemosensitive coefficients depend also on the concentration of the treatment. These modified coefficients will lead to changes in the size of the aggregates as shown in the numerical experiments in Section 4.5. The evolution of the concentration c is the same as in (4.7) where, in this case, the chemoattractant is produced by the new population w. Including proliferation also in this second part allows us to assess the effect of the treatment in the population at earlier times, while the population of cells is still growing and aggregates are still forming.

Different initial conditions for the cell density and concentration of the treatment will be considered as described in Section 4.5. Each initial condition for **P2** will correspond to a density profile solution of the system **P1** at a given time (solution of (4.16)), *i.e.* $w(\mathbf{x}, 0) = u(\mathbf{x}, T_1)$ for some given T_1. We will consider cases where the treatment is introduced on already formed and stable aggregates (steady state of (4.16)), but also cases where it is introduced at earlier times (during the formation of the aggregates, see Section 4.5). The initial condition for the concentration of the drug is considered to be either homogeneously distributed in the simulation domain or introduced in the center.

Remark 30 *In both parts of the experiments, we assume that the crowding capacity $\bar{u}$ is larger than the carrying capacity $u_{\max}$. The carrying capacity is defined as the maximum population density beyond which there are not enough nutrients to support growth, while the crowding capacity describes the maximum density in an aggregate depending on the space available.*

4.4 Linear stability analysis and pattern formation

The system for part **P1** (4.7) without source term is well known in the literature as the volume-filling Keller-Segel model, for which emergence of patterns has been characterised and is now well documented. Pattern formation refers to the phenomenon by which, after varying a bifurcation parameter, the spatially homogeneous steady state loses stability and inhomogeneous solutions appear. In the following, we investigate in which parameter region we can expect instability of homogeneous solutions, corresponding to the formation of patterns. The linear stability analysis followed here is classical and follows the lines of [162, 201, 149].

We first recall the two systems associated with the dynamics described in parts **P1** and **P2** of the model. Using the fact that in the first part of the experiments, the squeezing probability is chosen to be $q_1(u) = 1 - \frac{u}{\bar{u}}$, we have

$$\begin{aligned} \partial_t u &= \nabla \cdot (d_1 \nabla u - \chi_u \phi_1(u) \nabla c) + ru\Big(1 - \frac{u}{u_{\max}}\Big) \ , \\ \partial_t c &= d_2 \Delta c + \alpha u - \beta c \ , \end{aligned} \tag{4.12}$$

where

$$D_1 = 1 \ , \quad \text{and} \quad \phi_1(u) = u\big(1 - \frac{u}{\bar{u}}\big) \ . \tag{4.13}$$

For part **P2**, when the squeezing probability function is given by (4.10), the system writes

$$\begin{aligned} \partial_t w &= \nabla \cdot (d_3 D_2(w,M)\nabla w - \chi_w \phi_2(w,M)\nabla c) + \tilde{r} w\Big(1 - \frac{w}{u_{\max}}\Big) \ , \\ \partial_t c &= d_2 \Delta c + \alpha w - \beta c \ , \\ \partial_t M &= d_4 \Delta M - \delta w \ , \end{aligned} \tag{4.14}$$

with diffusion and chemotactic coefficients given by

$$D_2(w,M) = 1 + (\gamma_M - 1)\Big(\frac{w}{\bar{u}}\Big)^{\gamma_M} \quad \text{and} \quad \phi_2(w,M) = w\Big(1 - \Big(\frac{w}{\bar{u}}\Big)^{\gamma_M}\Big) \ . \tag{4.15}$$

4.4.1 Dimensionless model

To get a deeper insight to the behaviour of the system we introduce the characteristic values of the physical quantities appearing in the models. Denoting by X and T the macroscopic units of space and time, respectively, such that $\bar{\mathbf{x}} = \frac{\mathbf{x}}{X}$, $\bar{t} = \frac{t}{T}$, then we choose

$$(\bar{\mathbf{x}}, \bar{t}) = \left(\sqrt{\frac{\beta}{d_2}}\mathbf{x}, \ \frac{\beta d_1}{d_2} t\right) \ .$$

Using these new variables, the dimensionless systems write for part **P1**

$$\begin{aligned} \partial_t u &= \nabla \cdot (\nabla u - A\phi_1(u)\nabla c) + r_0 u\left(1 - \frac{u}{u_{\max}}\right) \ , \\ \zeta \partial_t c &= \Delta c + u - c \ . \end{aligned} \tag{4.16}$$

Similarly, we obtain for **P2**

$$\begin{aligned} \theta \partial_t w &= \nabla \cdot (D_2(w,M)\nabla w - B\phi_2(w,M)\nabla c) + \tilde{r}_0 n\Big(1 - \frac{w}{u_{\max}}\Big) \ , \\ \zeta \partial_t c &= \Delta c + n - c \ , \\ m \partial_t M &= \Delta M - \delta_0 w \ , \end{aligned} \tag{4.17}$$

where

$$A = \frac{\chi_u}{d_1} \ , \ r_0 = \frac{d_2 r}{d_1 \beta} \ , \ \zeta = \frac{d_1}{d_2} \ , \ \theta = \frac{d_1}{d_3} \ , \ B = \frac{\chi_w}{d_3} \ , \ \tilde{r}_0 = \frac{d_2 \tilde{r}}{d_3 \beta} \ , \ m = \frac{d_3}{d_4} \ , \delta_0 = \frac{\delta}{d_1} \ . \tag{4.18}$$

The parameters ζ and m are assumed to be small since the chemoattractant and the chemotherapeutic treatment diffuse faster than the cells. On the other hand, $\theta \simeq 1$ since both population densities u and w are assumed to be diffusing at similar rates. In the following, we state the linear stability for both systems in separate sections.

4.4.2 First part: Formation of the aggregates

We first consider the system (4.16), which can be re-written in a more general form as

$$\begin{aligned} \partial_t u &= \nabla \cdot (\nabla u - A\phi_1(u)\nabla c) + f(u) \ , \\ \zeta \partial_t c &= \Delta c + g(u,c) \ , \end{aligned} \tag{4.19}$$

where $\phi_1(u)$ is given in (4.13), $f(u) = r_0 u(1 - \frac{u}{u_{\max}})$ and $g(u,c) = u - c$. This system is subject to uniformly distributed initial conditions and zero-flux boundary conditions as in (4.8).

The main result in this section is the following theorem, which gives the conditions for pattern formation for the system (4.19).

Theorem 31 *Consider $(u^*,\ c^*)$ a spatially homogeneous steady state. Then pattern formation for the system (4.19) with zero flux boundary conditions (4.8) and initial data (4.9) is observed if the following conditions are satisfied,*

$$\begin{gathered} f_u^* + \zeta^{-1} g_c^* < 0\ ,\ f_u^* g_c^* > 0\ ,\ \zeta^{-1} g_c^* + f_u^* - \zeta^{-1} g_u^* A\phi_1(u^*) > 0\ , \\ g_c^* + f_u^* + g_u^* A\phi_1(u^*) > 2\sqrt{f_u^* g_c^*}\ . \end{gathered} \tag{4.20}$$

The critical chemosensitivity is given by

$$A^c = \frac{2\sqrt{r_0} + 1 + r_0}{u_{\max}\left(1 - \frac{u_{\max}}{\bar{u}}\right)}\ , \tag{4.21}$$

and for $A > A^c$ patterns can be expected. The wavemodes k^2 are in the interval defined by

$$k_1^2 = \frac{-m - \sqrt{m^2 - 4 f_u^* g_c^*}}{2} < k^2 < k_2^2 = \frac{-m + \sqrt{m^2 - 4 f_u^* g_c^*}}{2}\ , \tag{4.22}$$

where $m = -(g_c^* + f_u^* + g_u^* A\phi_1(u^*))$.

Proof. See Appendix 4.B for the proof of this theorem.

4.4.3 Second part: Treatment

We now consider the system (4.17) which corresponds to the second part of the experiments, when the treatment is introduced. The parameter range where patterns are observed is summarised in the following theorem.

Theorem 32 *Consider (w^*, c^*, M^*) a spatially homogeneous steady state. Also, consider (4.17) with zero flux boundary conditions given by*

$$(d_3 D_2(w,M)\nabla w - \chi_w \phi_2(w,M)\nabla c)\cdot \eta = 0\ ,\quad d_2 \nabla c \cdot \eta = 0\ ,\quad d_4 \nabla M \cdot \eta = 0\ ,$$

and initial conditions

$$\begin{aligned} &w(\mathbf{x},0) = u(\mathbf{x},T_1)\ , \\ &c(\mathbf{x},0) \text{ of } \textbf{\textit{P2}} \text{ is equal to } c(\mathbf{x},T_1) \text{ from } \textbf{\textit{P1}}\ , \\ &M(\mathbf{x},0) = \begin{cases} \text{constant in } \Omega\ , \text{ or} \\ Ce^{\frac{(\mathbf{x}-\mathbf{x}_0)^2}{2\sigma^2}}\ , \end{cases} \end{aligned} \tag{4.23}$$

where C is the amplitude, $\mathbf{x}_0$ is the centre of the Gaussian and σ is the width. Then, the critical chemosensitivity is given by

$$B^c = \frac{2\sqrt{\bar{r}_0 D_2(u_{\max}, M_s)} + D_2(u_{\max}, M_s) + \tilde{r}_0}{u_{\max}\left(1 - \left(\frac{u_{\max}}{\bar{u}}\right)^{\gamma_{M_s}}\right)}\ ,$$

where

$$D_2(u_{\max}, M_s) = 1 + (\gamma_{M_s} - 1)\left(\frac{u_{\max}}{\bar{u}}\right)^{\gamma_{M_s}}\ .$$

Patterns can be expected if $B > B^c$ *and the wavemodes* k^2 *are in the interval defined by*

$$k_1^2 = \frac{-\bar{m} - \sqrt{\bar{m}^2 - 4D_2(u_{\max}, M_s)(f_w^* g_c^*)}}{2D_2(u_{\max}, M_s)} < k^2 < k_2^2 = \frac{-\bar{m} + \sqrt{\bar{m}^2 - 4D_2(u_{\max}, M_s)(f_w^* g_c^*)}}{2D_2(u_{\max}, M_s)} ,$$

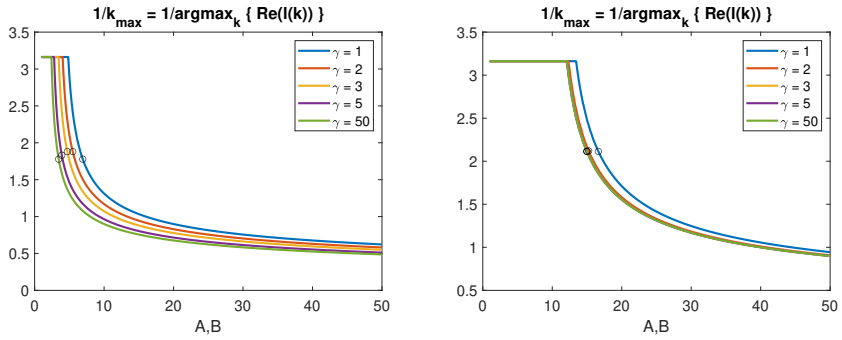
for $\bar{m} = -(D_2(u_{\max}, M_s)g_c^* + g_w^* B\phi_2(u_{\max}, M_s) + f_w^*)$.

Proof. The proof of this theorem can also be found in Appendix 4.B.

Remark 33 *For the case of* 2 *dimensions, we can rewrite the systems (4.16) and (4.17) using polar coordinates* (ρ, θ) *where we use the transformation* $x = \rho\sin\theta$, $y = \rho\cos\theta$ *and the Laplace operator is now given by* $\Delta_p = \frac{1}{R}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial}{\partial\rho}\right) + \rho^2\frac{\partial^2}{\partial\theta^2}$ *where* R *is the radius of the domain. The eigenvalue problem (4.40) is now written as* $-\Delta_p\psi_k = k^2\psi_k$ *with boundary conditions* $\partial\psi_k/\partial\rho = 0$ *at* $\rho = R$.

The eigenfunctions are obtained by separation of variables and are given by $\psi_k(x, y) = \mathcal{R}(\rho)\Theta(\theta)$. *Here* $\Theta(\theta) = e^{is\theta} = A\cos(s\theta) + B\sin(s\theta)$ *for some* $s \in \mathbb{Z}$. *The radial part* $\mathcal{R}(\rho)$ *is given in terms of Bessel functions* $\mathcal{R}(\rho) = \mathcal{J}_s(k\rho)$ *(see [177]) where* $k = \frac{c_{s,p}}{R}$ *and* $c_{s,p}$ *denotes the the pth zero derivative of* $\mathcal{J}_s$, *which is a first kind Bessel function of order* m. *Finally we can write* $\psi_k^{s,p}(\rho, \theta) = \mathcal{J}\left(\frac{c_{s,p}}{R}\rho\right)(A\cos(s\theta) + B\sin(s\theta))$.

The stability analysis reveals that several competing effects control the system's ability to create patterns (aggregates). The criteria obtained both in **P1** or **P2** show that the chemotactic sensitivity must be large enough to compensate the smoothing effect of the diffusion term and of the logistic growth. On the other hand, one can observe from the bifurcation formulae that the ratio $\frac{u_{\max}}{\bar{u}}$ (carrying capacity *vs* density of the tight packing state) plays an important role in the emergence of patterns: larger values lead to more aggregated states. These results show that the logistic growth term has an intrinsic smoothing property, *i.e.*, it tends to force the density to equate the carrying capacity, while the chemotactic term acts as a attractive force and creates zones of higher density (recall that $u_{\max} < \bar{u}$). The aggregates are an expression of a balance in between these two competing effects, which are completely characterised by the stability criterion.



(a) (b)

Figure 4.4 – Wavenumbers for different values of γ_M when (a) $r_0 = 0.1$, $u_{\max} = u_0 = 0.5$; and for (b) $r_0 = 0.05$, $u_{\max} = u_0 = 0.1$. Circles indicate the values of k_c.

In order to give more insights about the size of the emerging patterns, we show in Figure 4.4 the values of the inverse of the maximal wavenumber as function of the chemotactic sensitivity

(denoted by A for part **P1** and B for part **P2**), for different values of the exponent γ_M. We recall that $\gamma_M = 1$ in **P1** (blue curve) where cells act as rigid bodies and $\gamma_M > 1$ in the presence of the treatment, where cells behave as semi-elastic entities (**P2**). Figure 4.4a shows the results for growth rate $r_0 = 0.1$ and carrying capacity $u_{\max} = 0.5$, Figure 4.4b for $r_0 = 0.05$ (slower growth) and $u_{\max} = 0.1$ (lower carrying capacity). In both figures, the black circles indicate the critical values for the chemotactic sensitivity above which the system is unstable. Here, the tight packing density is set to $\bar{u} = 1$. The maximal wavenumber corresponds to the most unstable mode, *i.e.* the perturbed wave that will grow the fastest. Therefore, the inverse of this maximal wavenumber is directly related to the size of the emerging patterns. As one can observe, an increase in the chemotactic sensitivity parameter correlates with a decrease in the observed pattern size, suggesting that the aggregates are smaller: larger chemotactic sensitivity leads to more aggregated clusters. Moreover, the aggregate size also decreases as cells pass from rigid bodies to semi-elastic entities (when γ_M increases). This is due to the fact that for larger values of γ_M, cells are more easily deformed and can aggregate more efficiently than when they behave as rigid spheres.

When we increase the ratio $\frac{u_{\max}}{\bar{u}}$ (compare Figure 4.4a and 4.4b), we observe that the critical value of the chemosensitivity above which patterns are generated is larger than for smaller ratios $\frac{u_{\max}}{\bar{u}}$. These results highlight once again the smoothing effect of the logistic growth: When the cell tight packing density is unchanged, decreasing the carrying capacity of the environment enhances cell death in the aggregates formed by chemotaxis, where cells try to reach the tight packing state. In this case, the critical chemosensitivity value must be large enough to compensate for the cell death induced by the logistic growth. Moreover, we observe that larger ratios $\frac{u_{\max}}{\bar{u}}$ induce less influence of the parameter γ_M. The cell aggregation abilities are mainly driven by the chemosensitivity intensity and not so much by the cell mechanical properties for large values of $\frac{u_{\max}}{\bar{u}}$.

4.5 Numerical simulations

In addition to the analytic results obtained in Section 4.4, we present numerical simulations for the two problems (4.16) and (4.17). This allows us to investigate the behavior of the solution of the models for different scenarios and range of parameters. It is well-known that a standard discretization of the Keller-Segel models can lead to nonphysical solutions due to the convective term. Here, we focus on a numerical method that preserves the non-negativity of the cell density using the upwind finite element method described in [166] for the simulation of the Cahn-Hilliard equation.

The calculation of the chemotactic coefficient follows the lines of [12]. Indeed, the finite volume scheme proposed in [12] is identical to the numerical method presented in this paper in dimension one. However, in higher dimensions, and since we also use a finite element method, the numerical scheme presented in [12] differs from the one in this section as detailed in Appendix 4.C.

4.5.1 Biological relevance of the model parameters

Here, we comment on the choice of the model parameters that we will use for the numerical simulations and how they relate to experimental known data. In [89, 63], the proliferation rate for well oxygenated glioma cells in vitro r was shown to lie between 0.5 and 1 day^{-1}. As the proliferation rate relies significantly on the nutrient, also smaller value seems to be biologically admissible in real conditions and following [63] we choose $r = 0.4$ day^{-1} and $r = 0.8$ day^{-1}

(corresponding to the non-dimensionalised parameter $r_0 = 0.05$ and $r_0 = 0.1$). As hypoxia-inducible factors (HIF) are supposed to be responsible for the chemotaxis motion of GBM cells, we suppose that the diffusion coefficient d_2 and consumption rate β for the chemoattractant are linked to biological measurements of the oxygen diffusion in human brain which were estimated in [194, 89] to $d_2 = 86.4$ mm^2 day^{-1} and $\beta = 8640$ day^{-1}. However, we chose to consider slightly lower values $d_2 = 8.64$ mm^2 day^{-1} and $\beta = 864$ day^{-1} owing to the large variability of this parameter according to the type of tissue (see [190]). For such values and using the scaling of Section 4.4.1, one unit of time of our model corresponds to approximately 0.1 day and one unit of space is 0.1 mm. As we found no experimental data on the chemotactic coefficient χ_u of glioma cells in response to chemoattractant concentration, the choice of this parameter is driven by the stability analysis and we find that the interesting regimes are obtained for a dimensionless chemosensitivity in between 7 and 70, corresponding to a chemotactic coefficient $\chi_u \in [0.6, 6]$ mm^2 day^{-1}. Moreover, as no measurements for glioma cells diffusion coefficient are available in the literature, the parameter d_1 is arbitrarily chosen to be 100 times smaller than the chemoattratant diffusion speed and we choose $d_1 = d_3 = 0.086$ mm^2 day^{-1}, i.e the non-dimensionalised parameter $\zeta = \frac{d_1}{d_2} = 0.01$.

4.5.2 Numerical results for a one dimensional case

For all numerical computations we choose the packing capacity $\bar{u} = 1$. We consider different proliferation rates $r_0 = 0.1,\ 0.05$ and different initial conditions and carrying capacities $u_{\max} = u_0 = 0.1,\ 0.5$. The nondimensional parameters given in (4.18) are $\zeta = m = 0.01$ and $\theta = 1$ since we assume that the chemoattractant c and the treatment diffuse much faster than the cells, while the motility of the cells is not affected by the treatment, so $d_1 \approx d_3$. The initial condition for the cell density u_0 is assumed to be randomly distributed in space. Similarly, we can also define the initial chemoattractant concentration c_0.

In this section we start by solving the systems (4.16) and (4.17) on the interval $[0, L]$ with homogeneous non-flux boundary conditions using the method described in Appendix 4.C. In Appendix 4.D we investigate the effect of the size of the domain as well as the effect of the parameters A and B on the formation and evolution of patterns. Moreover, using (4.17) we study the effect of the treatment using the solution of (4.16) at the final time T_1 as initial condition. We explore the case when we introduce the treatment at earlier stages of the formation of the aggregates.

We consider two different scenarios for the evolution of the concentration of the treatment. First, we assume that the treatment diffuses very fast in the whole domain so that the concentration is homogeneous from time $T_2 = 0$. The other case we consider, which is closer to real experiments, starts with a high concentration of the drug in the centre of the domain and this concentration diffuses over time according to the third equation in (4.17).

Comparison with the linear stability analysis In order to quantify the aggregate sizes and compare it to the ones predicted by the stability analysis, we use the Fourier transform of the numerical solution and extract the frequency that corresponds to the maximal Fourier mode. For the sake of this analysis, periodic boundary conditions are therefore considered. To this aim, we compute the discrete Fourier transform $\mathcal{F}[u](\mathbf{x}, t) = \hat{u}(\lambda, t)$ and define

$$k_{\max} = \arg\max_{\lambda}(|\hat{u}(\lambda)|) = \arg\max_{\lambda}\left(\sqrt{\mathrm{Re}(\hat{u}(\lambda))^2 + \mathrm{Im}(\hat{u}(\lambda))^2}\right),$$

which corresponds to the frequency of the largest Fourier mode. The inverse $(k_{\max})^{-1}$ of this maximal frequency relates to the pattern size. This maximal frequency of the Fourier transform of the solution is expected to correspond to the maximal wavenumber predicted by the stability analysis. We show in Figure 4.5 the values of $(k_{\max})^{-1}$ computed from the numerical solution (blue dotted line) compared to the predictions of the stability analysis (red curve), as function of the chemosensitivity, for $\gamma_M = 1$ (left figure) and $\gamma_M = 5$ (right figure). As one can observe, we obtain a very good agreement between the numerical values and the predictions of the stability analysis, and we recover the critical value of the chemosensitivity parameter above which the system generates patterns.

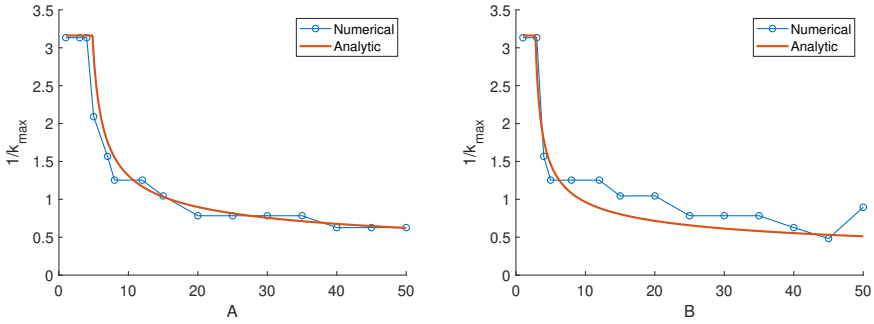


Figure 4.5 – Comparison of the wavelength obtained analytically, using (4.50) and (4.51), and numerically, using the Fourier transform of the solution for (a) $\gamma_M = 1$ and (b) $\gamma_M = 5$.

Introduction of the treatment on already-formed aggregates In this part, we aim to study the influence of the treatment on already formed aggregates. For this, we let the system run in **P1** (without treatment, $M = 0$, $\gamma_M = 1$) until time $T_1 = 200$ and introduce the treatment uniformly in the domain ($M = 1$, $\gamma_M = 5$).

In Figure 4.6 left and middle panels, we choose values of the chemosensitivity very close to the critical values corresponding to k_c, where the wavenumbers are very different for the cases $\gamma_M = 1$ and $\gamma_M = 5$ as we see in Figure 4.4a. In Figure 4.6, the blue curves describe the formation of aggregates for a time $T_1 = 200$ without the treatment, while the cells are proliferating with rate $r_0 = 0.1$. We consider two different scenarios when introducing the treatment: either cells stop proliferating (red curves), or they continue with the same rate as before $r_0 = 0.1$ (yellow curves).

When we introduce the treatment for values of A and B close to the instability threshold ($A = B = 7$, left figure), we observe that the aggregates become steeper and the density in each aggregate reaches the packing capacity $\bar{u} = 1$. This clearly leads to more compact aggregates as a result of the nonlinearity introduced in (4.11) by the parameter γ_M. The main physical difference between changing the parameter γ_M and changing the chemosensitivity coefficients A or B is the following. By changing A or B depending on the concentration of the treatment, we are enhancing aggregation over diffusion, essentially we are changing the motility of cells. By changing γ_M the motility, as well as the elastic properties of the cells in the aggregates are affected. When we introduce the treatment while cells keep proliferating, aggregates tend to merge together since the density is growing, as we see in Figure 4.6 middle panel.

It is noteworthy that for large values of the chemosensitivity parameter ($A = B = 70$, right panel in Figure 4.6), the treatment does not impact the aggregate dynamics. In this case, cell

aggregation is mainly driven by the chemotactic term and the cell mechanical properties have little influence. These observations are in agreement with the stability analysis, which shows that the parameter γ_M has more influence when the chemotactic sensitivity is close to the instability threshold.

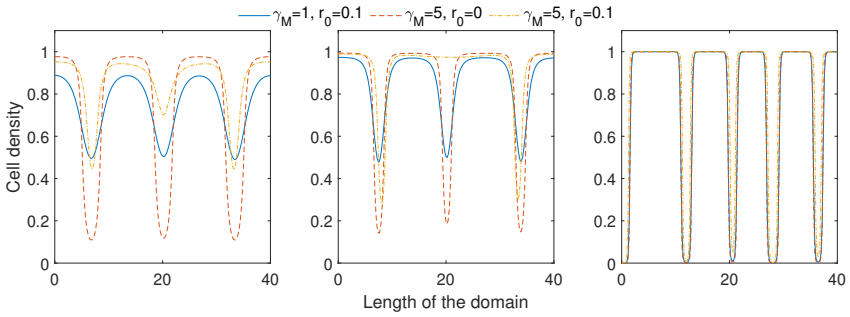


Figure 4.6 – Aggregation pattern when $A = B$ for $A = 7$ (left), $A = 12$ (middle) and $A = 70$ (right). Blue curve: solution at $T_1 = 200$ without treatment, red curves: solution at $T = T_1 + 200$ when the treatment is introduced uniformly with $\gamma_M = 5$ and proliferation is stopped, yellow curves: solution with the same parameters as the red curves when proliferation goes on with $r_0 = 0.1$.

Finally, by comparing the red and yellow curves of Figure 4.6, it is clear that cell proliferation has a major impact on the size of the aggregates. If the treatment has the double effect to stop proliferation as well as modifying the cell mechanical properties, cell aggregates will become very dense and well-separated, while merging aggregates are still observed if the treatment has the sole effect to change the cell mechanical properties. Here, we chose to introduce the treatment at time $T_1 = 200$, we now aim to study the effect of the treatment introduced at different times in the aggregation process.

Introduction of the treatment at different times Here, we consider the case when the treatment is introduced at different times in the aggregation process. As before, we let the system run in **P1** (without drug, $M = 0, \gamma_M = 1$) until time $t = T_1$ and introduce the treatment uniformly in the domain ($M = 1$, $\gamma_M = 5$). We consider the cases when the treatment has the ability to stop proliferation, and when the treatment only acts on the cell mechanical properties. In Figure 4.7, we show the results at time $T = T_1 + 200$ (red curve), when the treatment is introduced at times $T_1 = 50$ (left plots), $T_1 = 100$ (middle plots) and $T_1 = 300$ (right plots). Figures 4.7a show the results when the treatment stops proliferation, Figures 4.7b when the treatment only changes the mechanical properties of the cells. For each, the blue curves are the density profiles before introducing the treatment. As one can observe, in Figures 4.7a and 4.7b, introducing the treatment at different times of the aggregation process have a major impact on the size of the aggregated patterns formed at a latter time. Introducing the treatment at an earlier time ($T_1 = 50$, left figures) enables to obtain smaller aggregates compared to when the treatment is introduced on already formed aggregates ($T_1 = 300$, right figures). This effect is more visible when the treatment has the double effect of blocking cell proliferation and changing the elasticity (compare red curves in Figures 4.7a and 4.7b). In this case, the earlier the treatment is introduced, the smaller the aggregated patterns. When the treatment stops proliferation as well as the cell mechanical properties and is introduced at later times (right panel of Figure 4.7a) we recover the observation of the real systems, where the treatment induces a shrinking of the

aggregate and favors the formation of more compact cell structures. This effect is not observed when proliferation is active with the treatment, (right panel of Figure 4.7b) where aggregates are merging and they are larger than before the treatment introduction. This suggests indeed that the treatment has the double effect of blocking cell proliferation as well as changing the cell mechanical properties. The model suggests that introducing the treatment at earlier times of the tumour development could enable us to control the size and separation of the tumour aggregates.

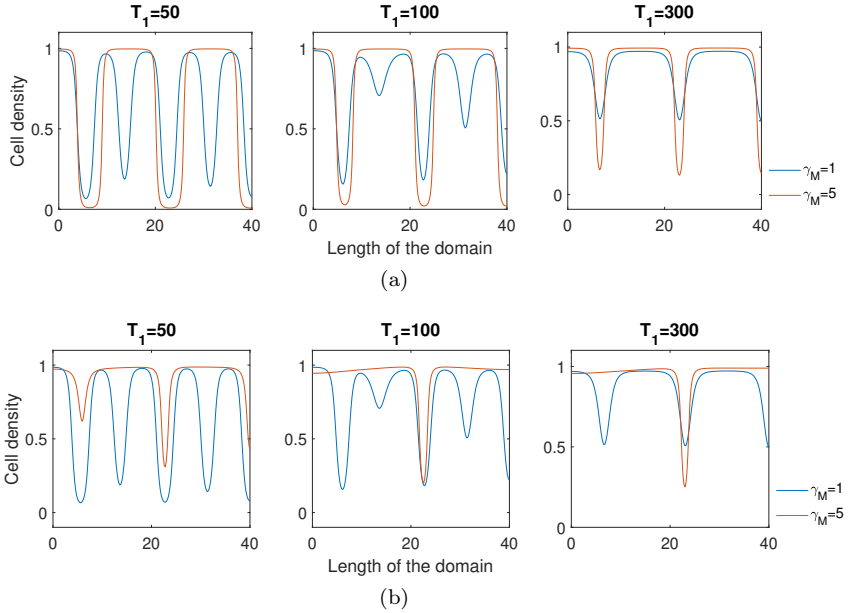


Figure 4.7 – Introduction of the the treatment at different times $T_1 = 50,\ 100,\ 300$ when $u_0 = 0.5$ and $A = B = 12$. The blue curves gives the initial condition for the part **P2** of the experiment with the treatment, represented by the red curve. (a) Without proliferation. (b) With proliferation, $r_0 = 0.1$. The red curves are at $T = T_1 + 200$.

Introduction of the treatment in the middle of the domain Finally, we aim to study the case when the treatment is introduced in the center of the domain and diffuses in the environment. Here we assume that the treatment is not consumed or escapes the domain, therefore $\delta_0 = 0$ in (4.17). In Figure 4.8, we show the density profiles of the solution before introducing the treatment (blue curves), and when the treatment is present (red curves), at times $T_2 = 0$ (left), $T_2 = 5$ (center) and $T_2 = 30$ (right). The distribution of the treatment follows a Gaussian of the form $M(\mathbf{x}, 0) = Ce^{\frac{(\mathbf{x}-\mathbf{x}_0)^2}{2\sigma^2}}$, where $C = 40$ is the amplitude, $\mathbf{x}_0$ is the center of the Gaussian and $\sigma = 0.5$ describes the spread. As one can observe, the large concentration of the treatment in the middle immediately sharpens the interface between already-formed aggregates, and favors the separation of the cell clusters. As the treatment diffuses (middle figure), the clusters interfaces which sense a high concentration of the treatment sharpen, creating denser and well-separated cell clusters.

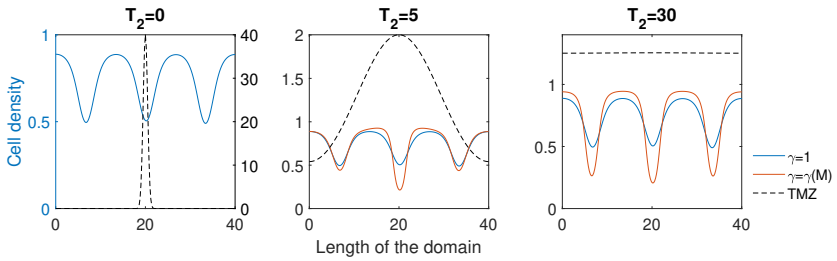


Figure 4.8 – Evolution of the pattern when the treatment is introduced in the center of the domain with amplitude 40. The blue curves correspond to the solution without treatment at $T_1 = 200$ and the red curves are the solution with the treatment at $T = T_1 + 5$ and $T = T_1 + 30$.

4.5.3 Numerical results for a two dimensional case

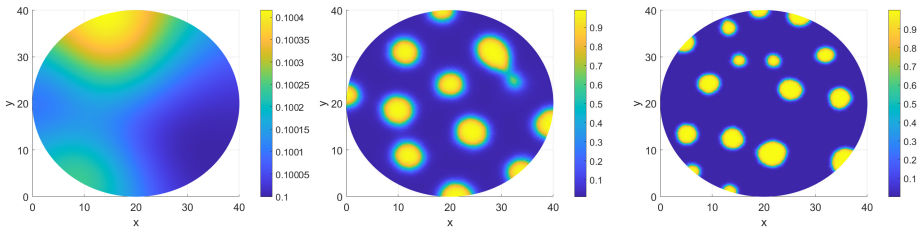
For the 2D simulations we consider that Ω is a disk of radius R which can be defined as $\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < R^2\}$ where the boundary is given by $\partial\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = R^2\}$. The proliferation rate is chosen to be $r_0 = 0.05$ and the initial homogeneous density as well as the carrying capacity are set to $u_{\max} = u_0 = 0.1$ or $u_{\max} = u_0 = 0.5$. The other parameters can be found at the beginning of Section 4.5.2.

In Figure 4.9 we show the formation of aggregates for different values of A without the treatment, for $u_0 = 0.1$ and $r_0 = 0.05$. We observe that for $A = 10$ (Figure 4.9a) we do not have patterns, in agreement with the analytic results obtained in Figure 4.7b since this value of A is less than $A^c \approx 16.7$. As we increase the chemosensitivity parameter, the aggregates become more compact. From Figure 4.9b we observe the phenomena of two aggregates merging together, analogous to the one dimensional results in Figure 4.15. As expected, by changing the carrying capacity and the initial density of cells to $u_0 = u_{\max} = 0.5$, the patterns change shape. We observe a transition from spot-like patterns in Figure 4.9 to maze-like structures in Figure 4.10. This behaviour has been widely studied experimentally [161], numerically [145, 150] and more recently, also including a volume-filling approach [162].

Analogous to the one dimensional case, we consider two different initial conditions for the treatment: (i) we first include the treatment uniformly in the domain with $M = 5$, and (ii) we introduce the treatment in the middle of the domain and let it diffuses in space. In these simulations, the treatment is supposed to block proliferation as well as changing the cell mechanical properties.

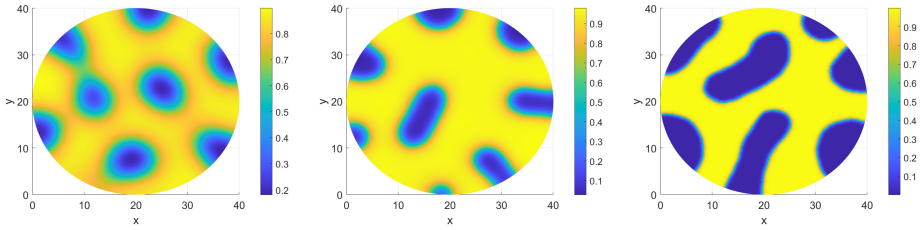
In order to compare the change in size of the aggregates before and after the treatment, we compute the difference between the solution of the first part of the experiments $u(\mathbf{x}, T_1)$ coming from (4.16), when the aggregates are formed (at $T_1 = 200$), and the solution $w(\mathbf{x}, t)$ of (4.17), once the treatment has been inserted (at time $T = 200 + T_2$). In Figure 4.11, we show the results for different values of the chemosensitivity parameter $A = B$, when the aggregates have been exposed to a uniform distribution of the treatment for a time $T_2 = 30$. We notice that the aggregates without the treatment are bigger and therefore the treatment induces a shrinking in the size of the pattern, as described in Section 4.3. When we compare the results in Figure 4.11a and 4.11b for different values of A and B we observe that the effect of the treatment is stronger when the value of the chemosensitivity parameter B is closer to its critical value (see Figure 4.4b). This is in accordance with the results of the stability analysis. The cell mechanical properties (controlled by the parameter γ_M) have less influence on the cell cluster sizes when the chemosensitivity parameter $A = B$ is increased (see Figure 4.4b).

We now study the case when the treatment is introduced in the middle of the domain and



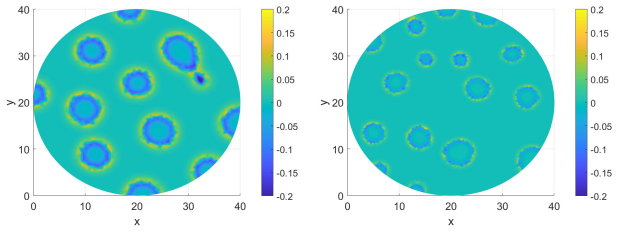
(a) (b) (c)

Figure 4.9 – Formation of aggregates at $T_1 = 200$ when $u_0 = 0.1$, $r_0 = 0.05$ and (a) $A = 10$, (b) $A = 20$ and (c) $A = 70$.



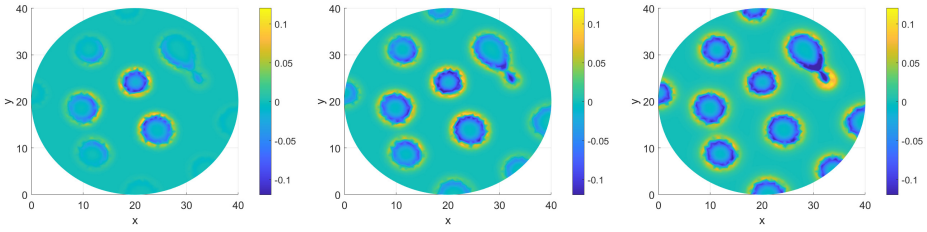
(a) (b) (c)

Figure 4.10 – Formation of aggregates at $T_1 = 200$ without the treatment and when $u_{\max} = u_0 = 0.5$, $r_0 = 0.05$ and (a) $A = 7$, (b) $A = 12$ and (c) $A = 50$.



(a) (b)

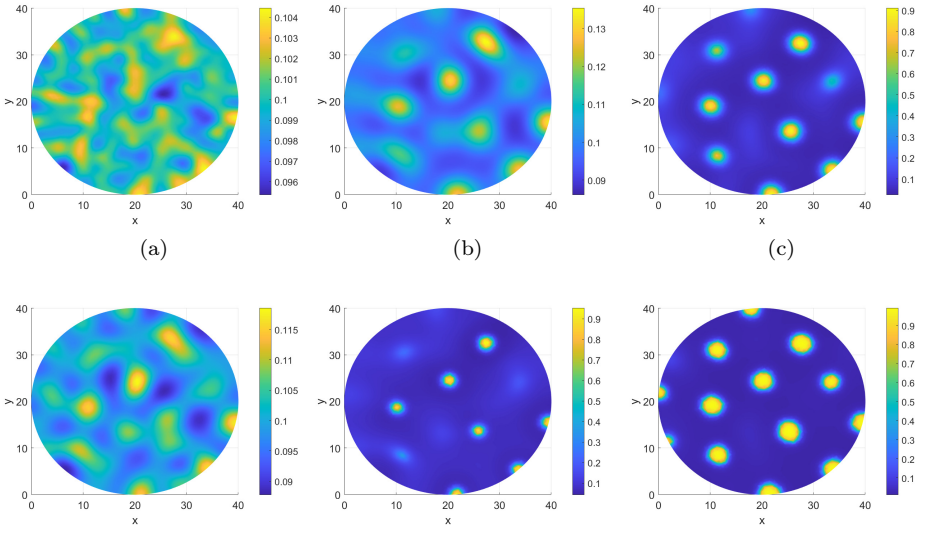
Figure 4.11 – Difference between the solutions $u(\mathbf{x}, 200) - w(\mathbf{x}, 230)$ when (a) $A = B = 20$ and (b) $A = B = 70$. Here $\gamma_M = 11$, $r_0 = 0.05$ and $u_{\max} = u_0 = 0.1$.



(a) (b) (c)

Figure 4.12 – Difference between the solutions when the initial concentration of the treatment is a Gaussian function centered in the domain. (a) $T = 210$, (b) $T = 230$ (c) $T = 300$ for $T_1 = 200$, $r_0 = 0.05$ and $u_0 = 0.1$.

diffuses in the environment. To this aim, the initial concentration of the treatment is supposed to be a Gaussian function with width 5 centered in the middle of the domain. We assume that the treatment is not consumed by the cells in the time scales we are interested in, and choose $\delta_0 = 0$. In Figure 4.12 we show the evolution of the difference between the two solutions $u(\mathbf{x}, T_1) - w(\mathbf{x}, t)$, where $T_1 = 200$ is the time at which the treatment is introduced and T is the duration of the treatment. We explore different times $T_2 = 10,\ 30,\ 100$. For short times, the effect of the treatment is only noticed by the aggregates at the center of the domain and therefore the difference between the two solutions close to the boundaries is zero. As time increases, the concentration of the treatment reaches the whole domain as is observed in Figure 4.12c.



(a) (b) (c)

(d) (e) (f)

Figure 4.13 – Effect of the treatment at different times while the aggregates are forming. Top row: Cell aggregates for $A = 20$ without treatment at (a) $T_1 = 30$, (b) $T_1 = 50$ and (c) $T_1 = 100$. Bottom row: Cell aggregates at time $T = T_1 + 200$, when the treatment has been introduced at times (d) $T_1 = 30$, (e) $T_1 = 50$ and (f) $T_1 = 100$.

Finally, we also study the effect of the treatment at earlier stages of the formation of the aggregates. Figure 4.13 shows the different patterns obtained during the formation of the aggregates at different times, top row, and the corresponding effect of the treatment, bottom row. For example, introducing the treatment at $T_1 = 50$ leads to a significant reduction of the size of the pattern with a reasonably low concentration of the treatment. Identifying this specific time in real patients could make the treatment much more effective and reduce the spread of the cancer cells.

4.6 Discussion of results and perspectives

In this paper we propose a mechanism for the effect of certain treatments on tumours formed by a chemotaxis type self-organization phenomenon. Inspired by the experiments concerning the action of TMZ on Glioblastoma cells mentioned in Section 4.2, we considered the particular case of treatments that do not act as cytotoxic agents but rather induce stress in the environment, which may induce changes in the mechanical properties of individual tumour cells by making them pass from rigid bodies to semi-elastic entities. We explored two scenarios: a first one where only cell's plasticity is impacted by the treatment, and a second one where the treatment has a double effect of preventing cell proliferation as well as changing cell mechanics.

Under these hypotheses, we obtained a modified version of the Keller-Segel model, known as the nonlinear volume-filling approach for cell motion, where the cell mechanical properties are taken into account in the form of the so-called squeezing probability. In the nonlinear volume-filling Keller-Segel model, this squeezing probability function could be related to the amplitude of the transport term towards zones of high chemoattractant density (chemosensitivity), as well as with the (nonlinear) diffusion coefficient.

By performing a linear stability analysis, we are able to detect and characterise the parameter ranges for which the homogeneous distribution is unstable, *i.e.* the ranges for which patterns appear as the result of the dynamics. We show that the emergence of patterns without treatment (*i.e.* when cells act as rigid bodies) is driven by a fine interplay between the chemotactic sensitivity, which tends to aggregate the cells, and the diffusion, together with the logistic growth, which tend to smoothen the solution. We are able to compute the critical chemosensitivity value above which the system self-organizes into aggregates, and characterise the size of the aggregates as a function of the model parameters.

Under treatment, *i.e.* when cells behave as semi-elastic entities, we show that the critical value of the chemosensitivity above which patterns emerge is smaller than that in control cultures, showing that as cells become more elastic, they tend to aggregate more easily than when they behave as rigid entities.

We are able to completely characterise the size of the patterns and show that semi-elastic cells create smaller aggregates than rigid entities for the same value of the chemosensitivity. These results suggest that the mechanical properties of individual cells have a huge impact on the shape and size of the aggregated patterns at the population level.

Moreover, we show that the ratio between the tight packing cell density and the carrying capacity of the TME plays a major role in the size and shape of the obtained patterns. For large values of this ratio, the aggregation abilities of the system are essentially driven by the chemotactic transport term while the individual cell mechanical behaviour has little impact on the shape and size of the patterns. However, for smaller values of this ratio, *i.e.* when the tight packing density is closer to the carrying capacity of the environment, cell mechanics has a huge influence on the behavior of the population.

These results are confirmed by numerical experiments in 1D and 2D for which, given an initial

perturbation of the homogeneous cell distribution, we observe the emergence of cell aggregates and we recover the critical values of the chemosensitivity predicted by the stability analysis. We obtain a very good correspondence between the simulations and the theoretical predictions, for the appearance of patterns as well as their size.

By performing simulations of the whole system, we recover the experimental observations: introducing the treatment (TMZ in the experiments mentioned in Section 4.2) on already-formed aggregates leads to the quick formation, of more compact patterns. As the treatment diffuses in the domain (changing locally the cell mechanical properties as it goes), it sharpens the border of the cell aggregates and leads to denser and well-separated clusters.

While the border sharpening of the clusters is independent on whether proliferation is activated or not during treatment, the shrinking of the clusters is more clear when the treatment has the double effect of changing the cell mechanical properties as well as blocking cell proliferation. Indeed, when proliferation is still active in the presence of the treatment, we observe the merging of existing clusters and this results in aggregates being larger than before treatment. These results suggest a possible mechanism for the shrinking of the aggregates observed under the experimental conditions described in Section 4.2: TMZ might not only stop cell proliferation, but might also generate a stress in the environment to which cells respond to by changing their mechanical state.

While alterations of mechanical properties, around or inside the tumour, are common in solid tumours including GBM, the question of the nature and the regulation of cancer cells through mechano-sensitive pathways are largely unknown. Recently, in a Drosophila model, glioma progression has been associated to a regulatory loop mediated by the mechano-sensitive ion channel Piezo1 and tissue stiffness [57]. A direct perspective of these works consists in verifying the potential mechano-sensing effect of TMZ proposed in the present model, through direct measures in real systems by studying the mechanical properties of individual GBM cells which have been exposed to TMZ treatment. The targeting of mechano-sensitive pathways after TMZ treatment may provide new therapeutic angles in GBM and in more general settings.

Moreover, it would be interesting to identify other clinical settings where the effects of the treatment are similar to those of TMZ in GBM, and to check if the effects are due to changes in the tumor cell properties corresponding to the general hypothesis of the model constructed in this work.

In the future, better quantitative comparison with experiments will allow for systematic choice of parameters and validation of the mechanisms we propose here. From a biological point of view, a natural sequel of this work consists in studying the coupled effect of TMZ and irradiation. Indeed, even if TMZ alone seems not to suffice to decrease the tumour mass, the coupling of TMZ treatment with irradiation has been shown to have more efficient effects than irradiation alone [186, 27, 169]. It would also be interesting to introduce a second treatment with cytotoxic effects in this model, to study whether the mechanical changes of individual cells induced by TMZ could explain the better response of the system to irradiation treatments.

4.A Derivation of the general model

The modified Keller-Segel system (4.7), including the squeezing probability (4.3) was derived in one dimension in [162, 201] from a continuous-time and discrete-space random walk. In this appendix we give a more general (and formal) derivation in $\mathbb{R}^n$ starting from a kinetic equation, analogous to [160, 176]. For simplicity, we do not include cell proliferation in this derivation.

Let us consider a mesoscopic density $h(\mathbf{x}, t, \mathbf{v})$, where $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{v} \in V = \{\mathbf{x} \in \mathbb{R}^n : |\mathbf{x}| = 1\}$, which evolution is described by the following kinetic equation

$$\partial_t h + \mathbf{v} \cdot \nabla h = -\xi(c)h(\mathbf{x}, t, \mathbf{v}) + q(\sigma(\mathbf{x}, t)) \int_V T(\mathbf{v}, \mathbf{v}', c)h(\mathbf{x}, t, \mathbf{v}') \, \mathrm{d}\mathbf{v}' \ . \tag{4.24}$$

Equation (4.24) describes a so-called velocity jump model where, individuals moving in an almost straight line, described by the left hand side, switch velocities after stopping. The right hand side of (4.24) describes the density of cells that are stopping with a frequency $\xi(c) = \int_V T(\mathbf{v}', \mathbf{v}, c) \, \mathrm{d}\mathbf{v}'$, where the operator $T(\mathbf{v}, \mathbf{v}', c)$ gives the probability of a velocity jump from $\mathbf{v}'$ to $\mathbf{v}$. Note that the cells change velocity also depending on the chemoattractant concentration $c(\mathbf{x}, t)$.

The second term in the right hand side represents the individuals that start a new trajectory with a different velocity $\mathbf{v}'$. Note that this term is multiplied by the squeezing probability q since the change into a new velocity is also determined by the probability of finding a neighbouring space available.

Now we introduce a parabolic scaling $(\mathbf{x}, t) \mapsto (\bar{\mathbf{x}}/\varepsilon, \bar{t}/\varepsilon^2)$ into (4.24), where the bar denotes the new variables and $\varepsilon \ll 1$. Letting $T(\mathbf{v}, \mathbf{v}', c)$ be a small perturbation of a random turning event, $T(\mathbf{v}, \mathbf{v}', c) = T_0 + \varepsilon T_0 T_1(\mathbf{v}', \nabla c)$, we have, after dropping the bars,

$$\begin{aligned} \varepsilon^2 \partial_t h^\varepsilon + \varepsilon \mathbf{v} \cdot \nabla h^\varepsilon = q|V|T_0 u^\varepsilon(\mathbf{x}, t) + \varepsilon q T_0 \int_V T_1(\mathbf{v}', \nabla c)h^\varepsilon(\mathbf{x}, t, \mathbf{v}') \, \mathrm{d}\mathbf{v}' \\ - h^\varepsilon(\mathbf{x}, t, \mathbf{v})T_0|V| - \varepsilon T_0 h^\varepsilon(\mathbf{x}, t, \mathbf{v})|V|T_1(\mathbf{v}, \nabla c) \ , \end{aligned} \tag{4.25}$$

where

$$u^\varepsilon(\mathbf{x}, t) = \frac{1}{|V|} \int_V h^\varepsilon(\mathbf{x}, t, \mathbf{v}) \, \mathrm{d}\mathbf{v} \ . \tag{4.26}$$

Dividing by ε we get

$$\begin{aligned} \varepsilon \partial_t h^\varepsilon + \mathbf{v} \cdot \nabla h^\varepsilon = \frac{T_0}{\varepsilon} \Big(q u^\varepsilon(\mathbf{x}, t)|V| - h^\varepsilon(\mathbf{x}, t, \mathbf{v})|V| \Big) \\ + T_0 \Big(q \int_V T_1(\mathbf{v}', \nabla c)h^\varepsilon(\mathbf{x}, t, \mathbf{v}') \, \mathrm{d}\mathbf{v}' - h^\varepsilon(\mathbf{x}, t, \mathbf{v})|V|T_1(\mathbf{v}, \nabla c) \Big) \ . \end{aligned} \tag{4.27}$$

The leading order terms in (4.27), when $\varepsilon \to 0$, give $h(\mathbf{x}, t, \mathbf{v}) = q u(\mathbf{x}, t)$. Integrating with respect to $\mathbf{v} \in V$ in (4.25) we obtain a macroscopic conservation equation

$$\partial_t u + \nabla \cdot j = 0 \ , \tag{4.28}$$

where $j^\varepsilon = \frac{1}{\varepsilon |V|} \int_V \mathbf{v} h^\varepsilon(\mathbf{x}, t, \mathbf{v}) \, \mathrm{d}\mathbf{v}$ is the mean direction of the cells.

Finally, we have to obtain j^ε in terms of the macroscopic density u^ε. For that, we multiply (4.25) by $\mathbf{v}$ and integrate in V to get

$$\varepsilon^2 |V| \partial_t j^\varepsilon + \nabla \cdot \int_V \mathbf{v} \otimes \mathbf{v} h^\varepsilon \, \mathrm{d}\mathbf{v} = -T_0 |V|^2 j^\varepsilon - |V| T_0 \int_V \mathbf{v} T_1(\mathbf{v}, \nabla c) h^\varepsilon \, \mathrm{d}\mathbf{v} \ . \tag{4.29}$$

Letting $h^\varepsilon(\mathbf{x},t,\mathbf{v}) = qu + \varepsilon h^\perp(\mathbf{x},t,\mathbf{v}) + \mathcal{O}(\varepsilon^2)$ we obtain, in the limit as $\varepsilon \to 0$,

$$j = \frac{-\int_V \mathbf{v}\otimes\mathbf{v}\,\mathrm{d}\mathbf{v}}{T_0|V|^2}\nabla(qu) - \frac{\int_V \mathbf{v}T_1(\mathbf{v},\nabla c)\,\mathrm{d}\mathbf{v}}{|V|}uq\ . \tag{4.30}$$

Substituting (4.30) into the conservation equation (4.28) we obtain

$$\partial_t u - \nabla\cdot\Big(d_1\nabla(qu) - \chi(c)uq\Big) = 0\ , \tag{4.31}$$

where

$$d_1 = \frac{\int_V \mathbf{v}\otimes\mathbf{v}\,\mathrm{d}\mathbf{v}}{T_0|V|^2} \quad\text{and}\quad \chi(c) = -\frac{\int_V \mathbf{v}T_1(\mathbf{v},\nabla c)\,\mathrm{d}\mathbf{v}}{|V|}\ . \tag{4.32}$$

Considering the squeezing probability function $q(u)$ in (4.3) and using the chain rule for differentiation, $\nabla(qu) = (q - q'u)\nabla u$ where $q' = \frac{\mathrm{d}q}{\mathrm{d}u}$ we finally have, assuming $\chi(c) = \chi_u\nabla c$,

$$\begin{aligned}\partial_t u &= \nabla\cdot(d_1(q-q'u)\nabla u - \chi_u uq\nabla c)\\ &= \nabla\cdot(d_1 D(u)\nabla u - \chi_u\phi(u)\nabla c)\ .\end{aligned} \tag{4.33}$$

Remark 34 *With the choice (4.3) for the squeezing probability, the diffusion of the cells is enhanced* $\left(\frac{\mathrm{d}D(u)}{\mathrm{d}u} > 0\right)$*. This means that the elastic collisions of the cells may increase, locally, the random motion component.*

4.B Stability analysis

First part of the experiments We first observe that the homogeneous distributions $u(\mathbf{x},t) = u^*$ and $c(\mathbf{x},t) = c^*$ are steady-states solutions of system (4.16) for u^* and c^* such that $f(u^*) = 0$ and $g(u^*,c^*) = 0$. In order to study their stability, we consider the system without spatial variations

$$\partial_t u = f(u)\ ,\quad \zeta\partial_t c = g(u,c)\ , \tag{4.34}$$

and linearize the solution at (u^*,c^*). We obtain

$$\partial_t\sigma = G\sigma\ ,\quad \text{where}\quad \sigma = \begin{pmatrix} u-u^* \\ c-c^*\end{pmatrix}\quad\text{and}\quad G = \begin{pmatrix} f_u^* & 0 \\ \zeta^{-1}g_u^* & \zeta^{-1}g_c^*\end{pmatrix}\ , \tag{4.35}$$

where the quantities f_u^*, g_u^* and g_c^* are the linearization slopes of f and g: $f_u^* = f'(u^*)$, $g_u^* = \partial_u g(u^*,c^*)$, $g_c^* = \partial_c g(u^*,c^*)$. The steady state is linearly stable if $\mathrm{tr}(G) < 0$ and $\det(G) > 0$, which imposes the following constraints on the kinetic functions $f(u)$ and $g(u,c)$,

$$f_u^* + \zeta^{-1}g_c^* < 0 \quad\text{and}\quad f_u^* g_c^* > 0\ . \tag{4.36}$$

Note that in our case, $f_u^* = -r_0$, $g_u^* = 1$, $g_c^* = -1$ so the conditions are satisfied.

We now go back to the full chemotactic system (4.19). In order to investigate the stability of the homogeneous steady-state, *i.e.* the ability of the system to create patterns, we introduce a small parameter $\varepsilon \ll 1$ and write

$$u = u^* + \varepsilon\tilde{u}(\mathbf{x},t)\ ,\quad c = c^* + \varepsilon\tilde{c}(\mathbf{x},t)\ . \tag{4.37}$$

We substitute (4.37) into (4.19) and, computing the first order terms with respect to ε and

neglecting higher order terms, the linearized system reads

$$\begin{aligned} \partial_t u &= \Delta u - A\phi_1(u^*)\Delta c + u f_u^* + c f_c^* \ , \\ \zeta \partial_t c &= \Delta c + u g_u^* + c g_c^* \ , \end{aligned} \tag{4.38}$$

where $\phi_1(u^*) = u^* q_1(u^*)$. We now look for perturbations of the form

$$u(\mathbf{x},t) = \sum_k a_k(t)\psi_k(\mathbf{x}) \quad \text{and} \quad c(\mathbf{x},t) = \sum_k b_k(t)\psi_k(\mathbf{x}) \ , \tag{4.39}$$

where $(\psi_k)_{k\geqslant 1}$ is an orthonormal basis of $L^2(\Omega)$ and satisfies the following spatial eigenvalue problem

$$-\Delta\psi_k = k^2\psi_k \ , \quad \frac{\partial \psi_k}{\partial \eta} = 0 \ . \tag{4.40}$$

Then, the linearized system (4.38) can be written as

$$\begin{aligned} \partial_t(a_k) &= -a_k k^2 + A\phi_1(u^*)b_k k^2 + a_k f_u^* + b_k f_c^* \ , \\ \zeta \partial_t(b_k) &= -b_k k^2 + a_k g_u^* + b_k g_c^* \ , \end{aligned} \tag{4.41}$$

where k is the spatial eigenfunction, also called the wavenumber and $1/k$ is proportional to the wavelength ω. In matrix form we can write (4.41) as $\partial_t X_k(t) = P_k(t)X_k(t)$ where

$$X_k = \begin{pmatrix} a_k \\ b_k \end{pmatrix} \ , \quad P_k = \begin{pmatrix} -k^2 + f_u^* & A\phi_1(u^*)k^2 + f_c^* \\ \zeta^{-1} g_u^* & \zeta^{-1}(-k^2 + g_c^*) \end{pmatrix} \ . \tag{4.42}$$

Remark 35 *Since the solutions of the eigenvalue problem (4.40) are simply sines and cosines, the "size" of various spatial patterns is measured by the wavelength of the trigonometric functions. For example, in one dimension when $0 < x < L$, $\psi \propto \cos(n\pi x/L)$ and the wavelength is $\omega = 1/k = L/n\pi$, where $n \in \mathbb{Z}$.*

If the matrix P_k has eigenvalues with positive real part, then the homogeneous steady state $(u^*,\ c^*)$ is unstable, resulting in pattern formation. The characteristic polynomial related to (4.42) is given by $\ell^2 + a(k^2)\ell + b(k^2) = 0$ where

$$a(k^2) = (1+\zeta^{-1})k^2 - (f_u^* + \zeta^{-1} g_c^*) \ , \tag{4.43}$$

$$b(k^2) = \zeta^{-1}k^4 - \zeta^{-1}(g_c^* + f_u^* + g_u^* A\phi_1(u^*))k^2 + \zeta^{-1} f_u^* g_c^* \ . \tag{4.44}$$

The eigenvalues ℓ determine the temporal growth of the eigenmodes, and we require $\mathcal{R}e(\ell(k^2)) > 0$ for the homogeneous steady state to be unstable. Note that we only look for the eigenmodes $k \neq 0$ since we already guaranteed that the steady state is stable in the absence of spatial perturbations, *i.e.* $\mathcal{R}e(\ell(k^2 = 0)) < 0$ in (4.36).

From the conditions (4.36), we know that $a(k^2) > 0$, hence the instability can only occur if $b(k^2) < 0$ for some k so that the characteristic polynomial has one positive and one negative root. This implies

$$k^4 - (g_c^* + f_u^* + g_u^* A\phi_1(u^*))k^2 + f_u^* g_c^* < 0 \ . \tag{4.45}$$

We also know from (4.36) that $f_u^* g_c^* > 0$, then a necessary (but not sufficient) condition for $b(k^2) < 0$ is

$$g_c^* + f_u^* + g_u^* A\phi_1(u^*) > 0 \ .$$

Remark 36 *The bifurcation between spatially stable and unstable modes occurs when the critical expression* $b_{\min}(k^2_{\min}) = 0$ *is satisfied.*

Moreover, to satisfy (4.45) the minimum $b_{\min}$ must be negative [149]. Differentiation with respect to k^2 in (4.44) leads to

$$b_{\min}(k^2_{\min}) = -\frac{(g^*_c + f^*_u + g^*_u A\phi(u^*))^2}{4} + f^*_u g^*_c \ . \tag{4.46}$$

Hence, the condition $b_{\min} < 0$ implies

$$g^*_c + f^*_u + g^*_u A\phi_1(u^*) > 2\sqrt{(f^*_u g^*_c)} \ . \tag{4.47}$$

To summarise, we have obtained the following conditions for the generation of spatial patterns for the chemotaxis system (4.19),

$$\begin{gathered} f^*_u + \zeta^{-1} g^*_c < 0 \ , \ f^*_u g^*_c > 0 \ , \ \zeta^{-1} g^*_c + f^*_u - \zeta^{-1} g^*_u A\phi_1(u^*) > 0 \ , \\ g^*_c + f^*_u + g^*_u A\phi_1(u^*) > 2\sqrt{f^*_u g^*_c} \ . \end{gathered} \tag{4.48}$$

From the analysis in this section, and using the particular forms of $\phi_1(u)$, f and g as in (4.19), it is easy to see that the spatially homogeneous steady states are $(0,0)$ and $(u_{\max}, u_{\max})$. We can check that $(0,0)$ is an unstable steady state, therefore we only work with $(u_{\max}, u_{\max})$ which, on the contrary, is stable. The first and second properties in (4.48) are immediately satisfied, *i.e.*, $-(r_0 + \zeta^{-1}) < 0$ and $r_0/\zeta > 0$, respectively. Finally, we have to check that the third and fourth conditions are satisfied as well. We have that

$$-1 - r_0 + A\phi_1(u^*) > 2\sqrt{r_0} \ . \tag{4.49}$$

Therefore, (4.49) is a necessary condition for pattern formation for the original system (4.12). Considering A as a bifurcation parameter, we can obtain a critical value A^c, so that we observe pattern formation if $A > A^c$. From (4.49) we get

$$A^c = \frac{2\sqrt{r_0} + 1 + r_0}{u_{\max}\left(1 - \frac{u_{\max}}{\bar{u}}\right)} \ . \tag{4.50}$$

The corresponding critical wavenumber k^2_c is obtained from (4.46) using (4.49) as follows,

$$k^2_c = \frac{g^*_c + f^*_u + g^*_u A^c \phi_1(u^*)}{2} = \sqrt{f^*_u g^*_c} = \sqrt{r_0} \ . \tag{4.51}$$

This means that, within the unstable range, $\mathcal{R}e(\ell(k^2)) > 0$ has a maximum wavenumber given by k^2_c. The range of linearly unstable modes $k^2_1 < k^2 < k^2_2$ is obtained from $b(k^2) = 0$,

$$k^2_1 = \frac{-m - \sqrt{m^2 - 4 f^*_u g^*_c}}{2} < k^2 < k^2_2 = \frac{-m + \sqrt{m^2 - 4 f^*_u g^*_c}}{2} \ , \tag{4.52}$$

where $m = -(g^*_c + f^*_u + g^*_u A\phi_1(u^*))$.

Second part of the experiments Following the same steps as before we linearise the system (4.17) to get

$$\begin{aligned} \partial_t w &= D_2(w^*, M^*)\Delta w - B\phi_2(w^*, M^*)\Delta c + w f_w^* \ , \\ \zeta\partial_t c &= \Delta c + w g_w^* + c g_c^* \ , \\ m\partial_t M &= \Delta M - \delta_0 w \ , \end{aligned} \tag{4.53}$$

where

$$D_2(w^*, M^*) = 1 + (\gamma_M^* - 1)\Big(\frac{w^*}{\bar{u}}\Big)^{\gamma_M^*} \ , \quad \phi_2(w^*, M^*) = w\Big(1 - \Big(\frac{w^*}{\bar{u}}\Big)^{\gamma_M^*}\Big) \ ,$$

and γ_M^* is given by (4.5) evaluated at M^*. As in (4.39), we let

$$w(\mathbf{x}, t) = \sum_k a_k(t)\psi_k(\mathbf{x}) \ , \quad c(\mathbf{x}, t) = \sum_k b_k(t)\psi_k(\mathbf{x}) \ , \quad M(\mathbf{x}, t) = \sum_k c_k(t)\psi_k(\mathbf{x}) \ , \tag{4.54}$$

where $\psi_k(\mathbf{x})$ satisfies (4.40) and we obtain a system $\partial_t X_k(t) = P_k(t)X_k(t)$ where

$$X_k = \begin{pmatrix} a_k \\ b_k \\ c_k \end{pmatrix} \ , \quad P_k = \begin{pmatrix} -D_2(w^*, M^*)k^2 + f_w^* & B\phi_2(w^*, M^*)k^2 & 0 \\ \frac{1}{\zeta}g_w^* & \frac{1}{\zeta}(-k^2 + g_c^*) & 0 \\ -\frac{\delta_0}{m} & 0 & -\frac{1}{m}k^2 \end{pmatrix} \ . \tag{4.55}$$

Similar to the previous section, the characteristic polynomial is given by $a(k^2)\ell^3 + b(k^2)\ell^2 + c(k^2)\ell + d(k^2) = 0$ where $a(k^2) = -1$ and

$$b(k^2) = -\Big(D_2 + \frac{1}{\zeta} + \frac{1}{m}\Big)k^2 + f_w^* + \frac{g_c^*}{\zeta} \ , \tag{4.56}$$

$$\begin{aligned} c(k^2) = &-\Big(\frac{D_2}{\zeta} + \frac{D_2}{m} + \frac{1}{m\zeta}\Big)k^4 + \Big(\frac{g_c^* D_2}{\zeta} + \frac{g_c^*}{\zeta m} + \frac{f_w^*}{\zeta} + \frac{f_w^*}{m} + \frac{g_w^* B\phi_2}{\zeta}\Big)k^2 \\ &- \frac{f_w^* g_c^*}{\zeta} \ , \end{aligned} \tag{4.57}$$

$$d(k^2) = -\frac{D_2}{m\zeta}k^6 + \frac{1}{m\zeta}(D_2 g_c^* + g_w^* B\phi_2 + f_w^*)k^4 + \frac{1}{m\zeta}(-f_w^* g_c^*)k^2 \ . \tag{4.58}$$

In general, the stability analysis for this cubic polynomial will require the Ruth–Hurwitz stability criterion [119] which states that the steady state is unstable if the coefficients of $a(k^2)\ell^3 + b(k^2)\ell^2 + c(k^2)\ell + d(k^2) = 0$ satisfy the condition

$$\frac{1}{(a(k^2))^2}\big(b(k^2)c(k^2) - a(k^2)d(k^2)\big) < 0 \ .$$

However, from (4.55) we observe that one of the eigenvalues of the matrix P_k is given by $\ell_1 = \frac{-k^2}{m} < 0$. The remaining two eigenvalues can be computed from the upper-left matrix

$$\begin{pmatrix} -D_2(w^*, M^*)k^2 + f_w^* & B\phi_2(w^*, M^*)k^2 \\ \frac{1}{\zeta}g_w^* & \frac{1}{\zeta}(-k^2 + g_c^*) \end{pmatrix} \ , \tag{4.59}$$

following the same analysis as for the case without TMZ.

The characteristic polynomial $\ell^2 + \bar{a}(k^2)\ell + \bar{b}(k^2) = 0$ related to (4.59) has coefficients

$$\bar{a}(k^2) = \Big(D_2(w^*, M^*) + \frac{1}{\zeta}\Big)k^2 - f_w^* - \frac{g_c^*}{\zeta} \ , \tag{4.60}$$

$$\bar{b}(k^2) = \frac{D_2(w^*, M^*)}{\zeta}k^4 - \Big(\frac{D_2(w^*, M^*)g_c^*}{\zeta} + \frac{B\phi_2(w^*, M^*)g_n^*}{\zeta} + \frac{f_w^*}{\zeta}\Big)k^2 + \frac{f_w^* g_c^*}{\zeta} \ . \tag{4.61}$$

For the steady state to be unstable we require, as before, that $\mathcal{R}e(\ell(k^2)) > 0$. Since $\bar{a}(k^2) > 0$ the instability can only occur if $\bar{b}(k^2) < 0$. Computing $\frac{\mathrm{d}\bar{b}(k^2)}{\mathrm{d}k^2} = 0$ from (4.61) we obtain

$$k_{\min}^2 = \frac{D_2(w^*, M^*)g_c^* + g_w^* B\phi_2(w^*, M^*) + f_w^*}{2D_2(w^*, M^*)} \ . \tag{4.62}$$

Hence from the condition $\bar{b}_{\min}(k_{\min}^2) < 0$ we get

$$D_2(w^*, M^*)g_c^* + g_w^* B\phi_2(w^*, M^*) + f_w^* > \sqrt{4D_2(w^*, M^*)f_w^* g_c^*} \ . \tag{4.63}$$

The spatially homogeneous steady state is $(w^*,\ c^*,\ M^*) = (u_{\max},\ u_{\max},\ M_s)$, where $M_s = |\Omega|^{-1}\int_\Omega M(\mathbf{x}, 0)\,\mathrm{d}\mathbf{x}$. Therefore, from (4.63) we obtain a critical constant B^c so that for any $B > B^c$ we observe pattern formation. This critical constant is given by

$$B^c = \frac{2\sqrt{\bar{r}_0 D_2(u_{\max}, M_s)} + D_2(u_{\max}, M_s) + \tilde{r}_0}{u_{\max}\Big(1 - \Big(\frac{u_{\max}}{\bar{u}}\Big)^{\gamma_{M_s}}\Big)} \ , \tag{4.64}$$

where

$$D_2(u_{\max}, M_s) = 1 + (\gamma_{M_s} - 1)\Big(\frac{u_{\max}}{\bar{u}}\Big)^{\gamma_{M_s}} \ . \tag{4.65}$$

The corresponding critical wavemode is given by

$$k_c^2 = \frac{D_2(u_{\max}, M_s)g_c^* + g_w^* B^c\phi_2(u_{\max}, M_s) + f_w^*}{2D_2(u_{\max}, M_s)} = \frac{\sqrt{D_2(u_{\max}, M_s)(f_w^* g_c^*)}}{D_2(u_{\max}, M_s)} \ . \tag{4.66}$$

Finally, the unstable modes are $k^2 < k_c^2$, where from $\bar{b}(k^2) = 0$ we get

$$\begin{aligned} k_1^2 =& \frac{-\bar{m} - \sqrt{\bar{m}^2 - 4D_2(u_{\max}, M_s)(f_w^* g_c^*)}}{2D_2(u_{\max}, M_s)} < k^2 < k_2^2 \\ =& \frac{-\bar{m} + \sqrt{\bar{m}^2 - 4D_2(u_{\max}, M_s)(f_w^* g_c^*)}}{2D_2(u_{\max}, M_s)} \ , \end{aligned} \tag{4.67}$$

for $\bar{m} = -(D_2(u_{\max}, M_s)g_c^* + g_w^* B\phi_2(u_{\max}, M_s) + f_w^*)$.

4.C Description of the numerics

We denote $H^1(\Omega) = W^{1,2}(\Omega)$ the usual Sobolev space and the standard L^2 inner product is denoted by $(\cdot,\cdot)$. Let $\mathcal{T}^h$, $h > 0$, be a quasi-uniform mesh of the domain Ω consisting of N disjoint piecewise linear mesh elements K such that the discretized domain $\overline{\Omega}_h = \bigcup_{K\in\mathcal{T}^h}\overline{K}$.

Let $h_K := \mathrm{diam}(K)$ and $h = \max_K h_K$ and for $d = 2$, we choose linear triangular elements. In addition, we assume that the mesh is acute *i.e.* for $d = 2$, each angle of the triangles can not exceed $\frac{\pi}{2}$. We must stress that for $d = 2$ since the domain Ω is circular, a small error of approximation is committed using Ω_h. We consider the standard finite element space associated with $\mathcal{T}^h$

$$V^h := \{\chi \in C(\overline{\Omega}) : \chi\big|_K \in \mathbb{P}^1(K), \quad \forall K \in \mathcal{T}^h\} \subset H^1(\Omega) \ , \tag{4.68}$$

where $\mathbb{P}^1(K)$ denotes the space of first-order polynomials on K. Let N_h be the total number of nodes of $\mathcal{T}^h$, J_h the set of nodes and $\{x_j\}_{j=1,\dots,N_h}$ their coordinates. We call $\{\chi_j\}_{j=1,\dots,N_h}$ the standard Lagrangian basis functions associated with the spatial mesh. We define the standard Lagrangian interpolation operator by $\pi^h : C(\overline{\Omega}) \to V^h$. We also need the lumped scalar product to define the problem

$$(\eta_1, \eta_2)^h = \int_\Omega \pi^h \left(\eta_1(x)\eta_2(x)\right) \mathrm{d}x \equiv \sum_{x_j \in J_h} (1, \chi_j)\eta_1(x_j)\eta_2(x_j) \ , \quad \eta_1, \eta_2 \in C(\overline{\Omega}) \ .$$

We define the standard mass and stiffness finite element matrices as G and K, where

$$G_{ij} = \int_\Omega \chi_i \chi_j \, \mathrm{d}x, \quad \text{for } i, j = 1, \dots, N_h \ ,$$

$$K_{ij} = \int_\Omega \nabla\chi_i \nabla\chi_j \, \mathrm{d}x, \quad \text{for } i, j = 1, \dots, N_h \ .$$

In the following finite element approximation of the Keller-Segel problem, the mass matrix is lumped, *i.e.* the matrix becomes diagonal with each term being the row-sum of the corresponding row of the standard mass matrix,

$$G_{l,ii} := \sum_{j=1}^{N_h} G_{ij} \ , \quad \text{for } i = 1, \dots, N_h \ .$$

Given $N_T \in \mathbb{N}^\star$, let $\Delta t := T/N_T$ be the time-step where T is the time corresponding to the end of the simulation. Let $t_n := n\Delta t$, $n = 0, \dots, N_T - 1$ be the temporal mesh. We approximate the continuous time derivative by $\frac{\partial u_h}{\partial t} \approx \frac{u_h^{n+1} - u_h^n}{\Delta t}$. We define

$$u_h^n(x) := \sum_{j=1}^{N_h} u_j^n \chi_j(x) \ , \quad \text{and} \quad c_h^n(x) := \sum_{j=1}^{N_h} c_j^n \chi_j(x) \ ,$$

the finite element approximations of the cell density u and the concentration of chemoattractant c where $\{u_j^n\}_{j=1,\dots,N_h}$ and $\{c_j^n\}_{j=1,\dots,N_h}$ are unknowns and $\{\chi_j\}_{j=1,\dots,N_h}$ is the finite element basis. Then, the finite element problem associated with the system (4.16) reads as follows.

For each $n = 0, \dots, N_T - 1$, find $\{u_h^{n+1}, c_h^{n+1}\}$ in $S^h \times S^h$, such that for all $\chi \in S^h$

$$\left(\frac{u_h^{n+1} - u_h^n}{\Delta t}, \chi\right)^h + \left(D(u_h^n)\nabla u_h^{n+1}, \nabla\chi\right) = \left(A\left(\phi^{\text{upw}}(u_h^n)\right)\nabla c_h^n, \nabla\chi\right) + r_0\left(u_h^n\left(1 - \frac{u_h^n}{u_{\max}}\right), \chi\right)^h, \tag{4.69}$$

$$\zeta\left(\frac{c_h^{n+1} - c_h^n}{\Delta t}, \chi\right)^h = -\left(\nabla c_h^{n+1}, \nabla\chi\right) + \left(u_h^{n+1} - c_h^{n+1}, \chi\right)^h. \tag{4.70}$$

The finite element scheme associated with the system (4.17) including the effect of the treatment is the following

$$\theta\Big(\frac{w_h^{n+1} - w_h^n}{\Delta t}, \chi\Big)^h + \left(\overline{D}(w_h^n, M_h^n)\nabla w_h^{n+1}, \nabla\chi\right) = \left(B\left(\overline{\phi}_2^{\text{upw}}(w_h^n, M_h^n)\right)\nabla c_h^n, \nabla\chi\right) + \tilde{r}\left(w_h^n\Big(1 - \frac{w_h^n}{u_{\max}}\Big), \chi\right)^h, \tag{4.71}$$

$$\zeta\left(\frac{c_h^{n+1} - c_h^n}{\Delta t}, \chi\right)^h = -\left(\nabla c_h^{n+1}, \nabla\chi\right) + \left(w_h^{n+1} - c_h^{n+1}, \chi\right)^h, \tag{4.72}$$

$$m\left(\frac{M_h^{n+1} - M_h^n}{\Delta t}, \chi\right)^h = -\left(\nabla M_h^{n+1}, \nabla\chi\right) - \delta\left(M_h^{n+1}, \chi\right)^h. \tag{4.73}$$

In order to describe how the chemotactic coefficients ϕ^{uwp} and $\overline{\phi}_2^{\text{upw}}$ are computed, let us rewrite the discrete equation (4.69) into its matrix form

$$(G_L + \Delta t K_D)\underline{u}^{n+1} = G_L\underline{u}^n + \Delta t K_\phi \underline{c}^n + \Delta t G_L \underline{g}^n \ ,$$

where $\underline{u}^n$ and $\underline{c}^n$ are the vectors of coefficients which are the unknowns of the problem and $\underline{g}^n$ is a vector defined by

$$\left[\underline{g}^n\right]_i = \left(u_h^n\left(1 - \frac{u_h^n}{u_{\max}}\right)\right)(x_i) \ , \quad \text{for } i = 1, \dots, N_h \ .$$

We define the finite element matrices associated with the diffusion K_D and the advection K_ϕ

$$K_{D,ij} = \int_\Omega D(u_h^n)\nabla\chi_i\nabla\chi_j \ \mathrm{d}x \quad \text{for } i, j = 1, \dots, N_h \ , \tag{4.74}$$

$$K_{\phi,ij} = \int_\Omega \phi^{\text{upw}}\left(u_h^n(x_i), u_h^n(x_j)\right)\nabla\chi_i\nabla\chi_j \ \mathrm{d}x \quad \text{for } i, j = 1, \dots, N_h \ . \tag{4.75}$$

In (4.74), the integral is computed using Gauss quadrature to deal with a potential choice of nonlinear functional for $D(u_h^n)$. The exactness of the quadrature is obtained using the adequate number of Gauss points since $D(u_h^n)$ is a polynomial of order $\gamma(M) + 1$.

The chemotactic coefficient is computed using an upwing approach. For each element and

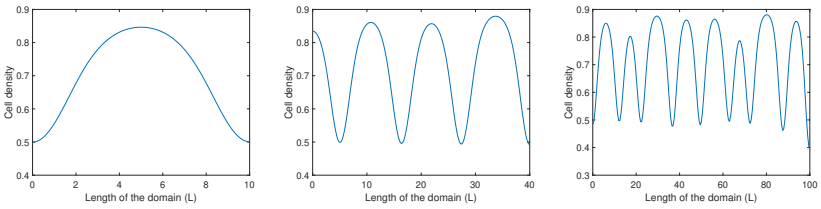
depending on the direction of the gradient of the chemoattractant we have

$$\phi^{\text{upw}}\left(u_h^n(x_i), u_h^n(x_j)\right) = \begin{cases} u_h^n(x_j)\left(1 - \left(\frac{u_h^n(x_i)}{\overline{u}}\right)\right), & \text{if} \quad c_h^n(x_j) - c_h^n(x_i) < 0, \\ u_h^n(x_i)\left(1 - \left(\frac{u_h^n(x_j)}{\overline{u}}\right)\right), & \text{otherwise.} \end{cases} \tag{4.76}$$

Therefore, the chemotactic coefficient is chosen as function of the sign of the difference of chemoattractant between nodes connected by an edge. The same method is applied to compute $\overline{\phi}_2^{\text{upw}}$ in (4.71). The property of non-negativity of the cell density satisfied by our numerical scheme can be proved using similar arguments as in [166].

4.D One dimensional numerical results

Influence of the domain size The unstable wavenumbers are discrete values, $k = n\pi/L$, that satisfy the relation (4.52) from Section 4.4.2. The wavemode n determines the number of aggregates depending on the length of the domain. For $A = 7$ and the parameters specified at the beginning of this section we have $0.25 < n\pi/L < 0.4$. As shown in Figure 4.14, as we increase the length of the domain, the number of aggregates also increases. When the domain is large, as in Figure 4.14c, we observe that some aggregates are merging together while others are emerging, *i.e.*, they are formed from a zone of low cell density. This process is called coarsening [162] and is not observed in a small domain such as in Figure 4.14a.



(a) (b) (c)

Figure 4.14 – Relationship between the wavenumber k and the length of the domain L for $A = 7$.

For the remaining simulations in this section we fix the length of the domain to $L = 40$, and all simulations are performed with carrying capacity $u_{\max} = 0.5$.

Comparison of the results with the stability analysis predictions Here, we study the influence of the chemosensitivity parameter A on the pattern dynamics and size of the aggregates, in the presence or absence of TMZ. In order to compare the solutions to the predictions of the stability analysis, the initial condition is a small perturbation around the homogeneous distribution $u_0 = 0.5$. Results of this section are obtained with a proliferation rate $r_0 = 0.1$. For such parameters, using the results of Section 4.4, the critical value of the chemosensitivity parameter without the treatment (in **P1**, when $M = 0$ and therefore $\gamma_M = 1$) is $A^c \approx 6.92$ and with the treatment uniformly distributed (for $\gamma_M = 5$) is $B^c \approx 3.9$.

In the first part of the experiment, *i.e.* without any treatment, we show in Figure 4.15 the formation of patterns at different times, for $A = 7$ (close to the instability threshold, Figure 4.15a), $A = 50$ (Figure 4.15b) and $A = 150$ (Figure 4.15c). We observe here the process of merging and emerging patterns through time.

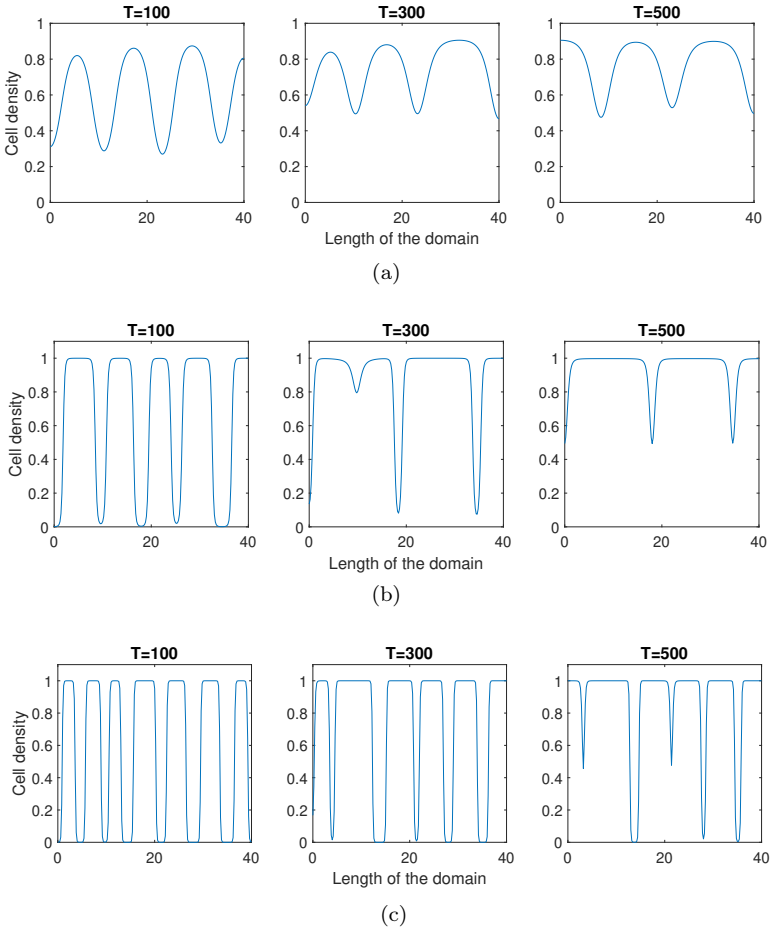
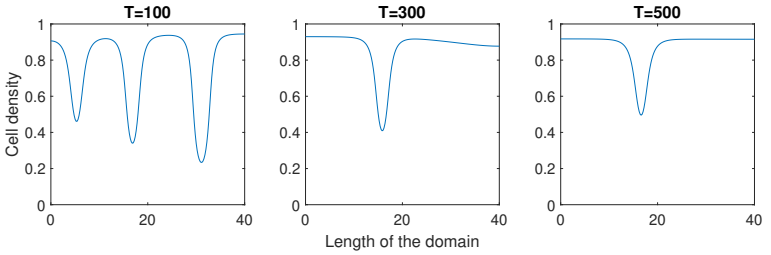


Figure 4.15 – Formation of the pattern without the treatment for (a) $A = 7$, (b) $A = 50$ and (c) $A = 150$ when $r_0 = 0.1$ and $u_{\max} = u_0 = 0.5$.

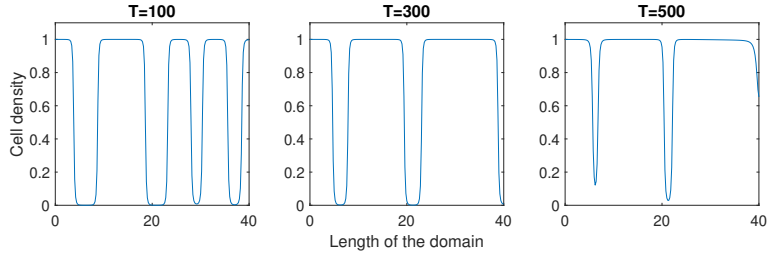
As predicted by the stability analysis, larger values of the chemotactic sensitivity A favor the emergence of smaller aggregates (compare Figures 4.15a and 4.15c for $A = 7$ and $A = 150$, respectively). This is due to the fact that for larger chemotactic sensitivity, cells are more attracted to zones of high concentration of chemoattractant. The creation of patterns instead of the expansion of a homogeneous cell distribution is due to an instability which results from a positive feedback loop between the production of the chemical by the cells on one hand, and their attraction to high density zones of this chemical on the other. The chemotactic sensitivity A must be large enough to trigger this instability, in order to compensate the competing effects of diffusion and of the logistic growth term, which on the contrary, tends to regulate the local cell density to the carrying capacity of the environment $u_{\max}$, and therefore induces cell death inside the aggregated patterns for which the density is above $u_{\max}$.

When the drug is introduced uniformly in the domain starting from a homogeneous distribution of cells (*i.e.* for $M = 1$, $\gamma_M = 5$) at time $t = 0$, we show in Figure 4.16 the formation of patterns at different times, for chemosensitivity $B = 5$ (close to the instability threshold, Figure 4.16a), $B = 30$ (Figure 4.16b) and $B = 150$ (Figure 4.16c). Note that in this case we also let cells to proliferate with rate $r_0 = 0.1$.

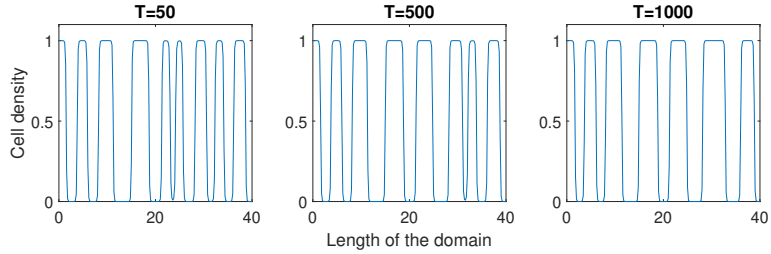
As one can see in Figure 4.16, we first observe again that increasing the chemosensitivity parameter B results in the formation of smaller cell aggregates (compare Figures 4.16a and 4.16b). Very close to the instability threshold (Figure 4.16a), the system converges quickly to one aggregate, while for larger values of B (Figure 4.16c) a large number of well-separated small aggregates arises. These clusters merge in time to form bigger clusters as for the case without the treatment. Moreover, comparing Figures. 4.15c and 4.16c, we clearly observe that varying the mechanical state of cells (*i.e.* passing from $\gamma_M = 1$ to $\gamma_M = 5$), leads to a change in the cell's aggregate size. When cells are more elastic, they tend to create smaller aggregates than when they behave as rigid spheres.



(a)



(b)



(c)

Figure 4.16 – Formation of the pattern with treatment ($M = 1, \gamma_M = 5$), included at $t = 0$ when $u_0 = 0.5$, $r_0 = 0.1$ for (a) $B = 5$, (b) $B = 30$ and (c) $B = 150$.

Chapter 5

Compressible Navier-Stokes-Cahn-Hilliard model for the modelling of tumor invasion in healthy tissue.

Abstract

We propose a compressible two-phase Navier-Stokes-Cahn-Hilliard model to represent the biological situation of two cell populations moving through a fibrous extracellular matrix. To study the emergence of irregular tumor borders during the invasion of the surrounding healthy tissues, we consider that one population proliferates and the two cell populations have different mechanical properties. Our generalized Navier-Stokes-Cahn-Hilliard model is consistent with the laws of mechanics and thermodynamics. From simple assumptions, we recover a simpler model that is similar to a system of two-porous medium equations. In this model, the two equations are coupled through a pressure term. Our study shows that when tumor cells move faster through the ECM than the cells constituting the healthy tissue, irregular tumor borders may emerge. In a future work, we will investigate the effect of cell-cell adhesion and cell viscosity on the emergence of these patterns by numerical simulations of different scenarios.

This chapter contains preliminary results of an ongoing work with Tommaso Lorenzi.

5.1 Introduction

We derive and study a mathematical models for the dynamics of two cell populations with different mechanical properties. We also assume that only one population is proliferating while the other is in homeostatic equilibrium. To consider the effect of the extracellular matrix, we assume that the cells are moving and exert adhesion effects on a mesh of hard fibers. We use a mixture model to represent the two cell populations, and we follow the evolution of their mass fraction. The mathematical representation of the mechanical interactions is given by a system composed of a Cahn-Hilliard type model coupled with a compressible Navier-Stokes equation for a viscous mixture. To consider the adhesion of cells on the ECM, we include a friction term in the equation for the velocity field. Altogether, our model can take into account possible differences in mechanical properties between the two cell populations.

We denote by ρ the mass density of the mixture, c the mass fraction of one of the cell phase and p the hydrostatic pressure. We pose our problem in a smooth bounded domain $\Omega \subset \mathbb{R}^d$,

$d = 1, 2, 3$. Our model reads

$$\begin{aligned}
\frac{\partial \rho}{\partial t} &= -\mathrm{div}\,(\rho \mathbf{v})\,\rho c G(p),\\
\rho \frac{\mathrm{D}c}{\mathrm{D}t} &= \mathrm{div}\,(b(c)\nabla \mu) + \rho c(1-c)G(p),\\
\rho \mu &= -\gamma \mathrm{div}\,(\rho \nabla c) + \rho \frac{\partial \psi_0}{\partial c},\\
\rho \frac{\mathrm{D}\mathbf{v}}{\mathrm{D}t} &= -\left[\nabla p + \gamma \mathrm{div}\,(\rho \nabla c \otimes \nabla c)\right] + \mathrm{div}\left(\nu(c)\left(\nabla \mathbf{v} + \nabla \mathbf{v}^T\right)\right)\\
&\qquad - \frac{2}{3}\nabla\left(\nu(c)\left(\mathrm{div}\,(\mathbf{v})\right)\right) - \kappa(c)\mathbf{v} - \rho c \mathbf{v} G(p),
\end{aligned}$$

supplemented by zero-flux Neumann boundary conditions

$$\frac{\partial \mu}{\partial \mathbf{n}} = \frac{\partial \mathbf{v}}{\partial \mathbf{n}} = \frac{\partial \rho c}{\partial \mathbf{n}} = \frac{\partial \rho}{\partial \mathbf{n}} = 0,$$

where $\mathbf{n}$ is the outward normal vector to the boundary $\partial\Omega$. Borrowing the terminology of the Cahn-Hilliard framework, we referred to μ as the chemical potential and ψ_0 as the homogeneous free energy that represents the mechanical interaction between cells. The phases of the fluid are separated by a smooth transition layer of width $\sqrt{\gamma}$. We also have that $b(c)$ is a degenerate mobility functional, $G(p)$ is a pressure-dependent growth function that represents the proliferation of cells, $\nu(c)$ is a viscosity coefficient and $\kappa(c)$ is a friction coefficient. These two latter functions are used to take into account the fact that for the two components of the mixture friction and mobility affects can be different.

Building upon [139], we aim at modelling the dynamics of two cells populations in a scenario whereby one of them is proliferating while the other is not, and the two populations move with different velocities. In [139], the authors proposed the model

$$\begin{cases} \frac{\partial \phi_1}{\partial t} - \kappa_1 \mathrm{div}\,(\phi_1 \nabla p) = \phi_1 G(p), \\ \frac{\partial \phi_2}{\partial t} - \kappa_2 \mathrm{div}\,(\phi_2 \nabla p) = 0, \end{cases} \tag{5.1}$$

where ϕ_1, ϕ_2 are the relative mass densities of each cell population, κ_1, κ_2 are the mobility coefficients that can be different, and $p = \rho^a$ is the pressure and $a \geq 1$ is a parameter controlling the stiffness of the pressure law. This model is used as a phenomenological representation of two cell populations with a possible application to the modelling of tumor protrusions in breast cancer. Numerical simulations of this model present two types of possible patterns depending on the values of the mobility coefficients. The initial condition that the authors considered represents a spherical core of cell of population 1 surrounded by cells of population 2. If $\kappa_1 < \kappa_2$, it appears a spherical wave of dividing cells pushing the surrounding non-dividing cells. However, if $\kappa_1 > \kappa_2$, as the proliferating population grows, finger-like instabilities emerge. These protrusions are comparable to the viscous figering patterns (i.e., Saffman-Taylor instabilities) seen in Hele-Shaw cells when a fluid is injected in a thin space contained between two parallel plates that contains a more viscous fluid [171].

The patterns observed in [139] are qualitatively similar to patterns of tumor cell growth observed in patients affected by breast cancer. Indeed, Figure 5.1 is a histological examination of an invasive breast tumor taken from [199]. An irregular invasive front is observed, and the authors in [139] suggested that a possible explanation of this phenomenon could be a contrast of motility between cancer cells and adypocytes. Figure 5.1 indicates also an other important

effect. A differences in size and number of adipocytes is observed close to the tumor protrusions. This seems to indicate that the physical properties of the adipocytes play an important role in the emergence of these irregularities of the tumor. Based on these observations, we are interested in understanding the key physical properties of tumors and cells of the healthy tissue that allow for the emergence of protrusions of the tumor inside the healthy tissue.

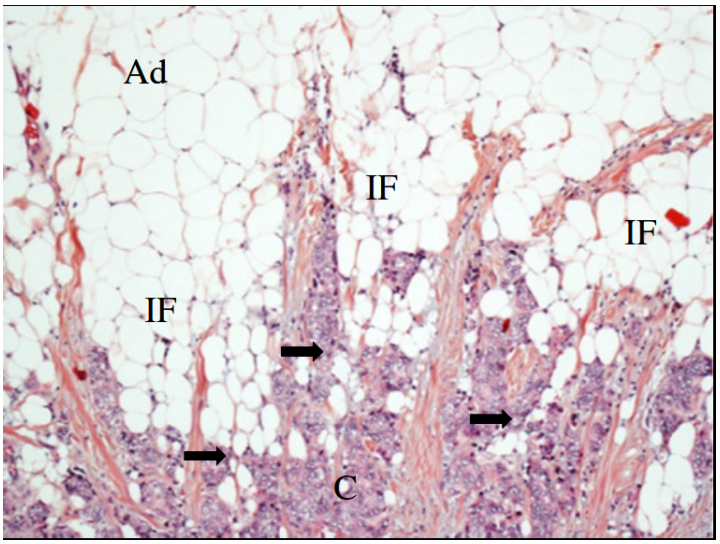


Figure 5.1 – Histological image of a breast tumor taken from [199]. Adipocytes are whites cells while tumor cells are purple. *IF* refers to the invasive front, and *C* to the tumor center. The dark arrows point to zones close to the protrusions where adipocytes have a smaller size and are less numerous. (with permission from the journal to reproduce the figure).

We aim to derive a general model that considers the effects of adhesion between cells and on the extra-cellular matrix, pressure, active motility, and viscosity. A complete description of all the different cell types present in and around the tumor would lead to a complex and lengthy system of equations. To simplify our mathematical representation, we focus only on two cell types: tumor and healthy cells. Our modeling approach relies on the theory of mixture for the representation of living tissue [46]. To highlight the crucial physical effects, we do not neglect any effects that can occur in the mixture: we keep inertia and compressibility. Altogether, our model is a thermodynamically consistent model of two-phase compressible fluid with a source term and friction term, i.e., a generalized Navier-Stokes-Cahn-Hilliard system. To understand the derivation of a multiphase fluid model with the theory of mixture within the framework provided by basic thermodynamics we refer the reader to [114].

This type of compressible Navier-stokes system coupled to a Cahn-Hilliard equation to take into account the effect of adhesion and repulsion forces between the component have been used in material sciences. Lowengrub and Truskinovsky [141] derived a model for compressible Cahn-Hilliard fluids consistent with thermodynamics. They investigated the assumption of incompressibility and showed that the velocity may not be solenoidal in that case, therefore referring to the mixture as a quasi-impressible fluid. In [1], Abels and Feireisl proved the existence of global weak solutions for a variant of the compressible model of Lowengrub and Truskinovsky [141]. Indeed, to use the method of Leray [131] and Lions [136], they used a thermodynamically consistent modification of the constitutive relation for the homogeneous free energy of the model.

In the present work, our model is a modification of the compressible Navier-Stokes-Cahn-

Hilliard model found in [141]. Indeed, we add a source term in the mass balance equation for the population 1 and a friction term in the equation for the velocity field. These two modifications are considered to take into account the effect of friction on the ECM and of proliferation of tumor cells. However, we show through a physically rigorous derivation that even with these modifications, the model remains consistent with thermodynamics. Altogether, this work aims at using a compressible two-phase Navier-Stokes-Cahn-Hilliard fluid to represent the progression of a tumor in a healthy tissue and understand what are the crucial mechanical properties of the tissues to observe the emergence of irregularities at the surface of growing tumors.

Numerous research pieces using the Cahn-Hilliard framework for the modeling of tumors can be found in the literature. Starting from the work of Khain and Sander [126], which used the Cahn-Hilliard equation for incompressible diphasic fluids as the continuous equivalent of a discrete model describing two populations of cells that can move, proliferate, and experience cell-cell adhesion, many authors started to model living tissues and in particular tumors as Cahn-Hilliard fluids. Indeed, Wise *et. al.* [203], and Frieboes *et. al.* [90] presented a Cahn-Hilliard framework for incompressible diphasic fluids to model tumor growth. In these three previous works, the potential representing attractive and repulsive forces (i.e., the homogeneous free energy) is a thermodynamically relevant double-well logarithmic potential function. Furthermore, the motilities of the two components inside the mixture are assumed equal and constant.

To better represent the movement of tumor cells using the Cahn-Hilliard framework, Chatelain *et. al.* [56, 55] and Agosti [8] used a logarithmic single-well potential to represent attractive and repulsive forces occurring between the components and a degenerate mobility function. This choice of potential is a phenomenological representation of the scenario where tumor cells are the only active components of the mixture, the other phase being only composed of inactive matter such as the extracellular fluid or non-active cells. This choice is also explained by the work of Byrne and Preziosi [46]. The two types of potential are investigated in our work to understand the effect of the attractive and repulsive forces exerted by the healthy cells. Altogether, the framework provided by the degenerate Cahn-Hilliard model with a single-well potential has shown very promising results and retrieves a qualitatively good agreement when compared to patient data [10, 9].

This type of Cahn-Hilliard model has been modified to include the representation of more physical effects. To this end, Garcke *et. al.* proposed a Cahn-Hilliard-Darcy model for two-phase incompressible fluids [98] (that has been extended to the case of multiphase fluids [97]) in which active and passive transport is considered as well as death and proliferation of cells. Active transport is characterized by movement due to attractive and repulsive forces between cells and chemotaxis due to nutrients that diffuse in the domain. The passive transport is given in the model by an advection term where a Darcy-type equation defines the velocity field. To take into account the effect of the fluid viscosity, this model has been modified using a Brinkman-type law to define the velocity field [71].

However, these two systems neglected an effect that seems to play a crucial role in our application, the compressibility of tissues. However, even if incompressibility is often assumed in the modeling, it is well-known that the compressibility of the healthy tissue due to tumor growth results in the alteration of the physical properties of the micro-environment [85, 153, 111, 152, 115]. Furthermore, it has been argued that compressibility also plays a role in the proliferation of tumor cells and their motility [132].

Therefore, to study the role of physical mechanisms during the invasion of a tumor in a healthy tissue, and to understand the emergence of irregular tumor borders, we propose a general two-phase Cahn-Hilliard fluid model. Our model includes the effects of compressibility of the tissues, viscosity, adhesion and crawling of cells on the extracellular matrix, attraction and repulsion between cells. The outline of this chapter is the following: in Section 5.2, we present

the derivation of our model and show its consistency with basic thermodynamics. This section gives the constitutive relations for the important functions of our system. Then, Section 5.3 presents the general assumptions on the important functions constituting the model. We also give some particular choices of functions for our application, and explain their biological relevancy. To understand the connection of our model to the system (5.1), we conduct in Section 5.4 formal asymptotic calculations assuming that the adhesion of cells on the ECM (i.e., the friction term in the equation for the velocity field) is the predominant effect. Section 5.5 presents the finite volume scheme that we will use in to construct numerical solutions to the model equations.

Remark. This study is undergoing. The numerical experiments are currently conducted. Many scenarios are tested to identify the crucial physical effects that produce the observed irregularities.

5.2 Derivation of the model

5.2.1 Notation and definitions

We formulate our problem in Eulerian coordinates and in a smooth bounded domain $\Omega \subset \mathbb{R}^d$ (where $d = \{1, 2, 3\}$ is the dimension). The balance laws that are derived in the following sections are in local form.

We have two cell populations in the model where ρ_1, ρ_2 are the relative densities of respectively population 1 and 2. Thus, ρ_i represents the mass of the population M_i per volume occupied by the i-th phase V_i i.e.

$$\rho_i = \frac{M_i}{V_i}.$$

Then, we define the volume fractions φ_1, φ_2 which are defined by the volume occupied by the i-th phase over the total volume of the mixture

$$\varphi_i = \frac{V_i}{V}.$$

Therefore, the mass density of population i which is the mass of population i in volume V is given by

$$\phi_i = \rho_i \varphi_i.$$

The total density of the mixture is then given by

$$\rho = \phi_1 + \phi_2.$$

We also introduce the mass fractions $c_i = M_i/M$ and we have the relations

$$\rho c_i = \phi_i, \quad \text{and} \quad c_1 = (1 - c_2). \tag{5.2}$$

We denote by p the hydrostatic pressure of the mixture and $\mathbf{v}_1, \mathbf{v}_2$ are the velocities of the different phases. We use a mass-average mixture velocity

$$\mathbf{v} = \frac{1}{\rho} \left(\phi_1 \mathbf{v}_1 + \phi_2 \mathbf{v}_2 \right). \tag{5.3}$$

We define the material derivative for a generic function g (scalar or vector-valued) by

$$\frac{\mathrm{D}g}{\mathrm{D}t} = \frac{\partial g}{\partial t} + \mathbf{v} \cdot \nabla g, \tag{5.4}$$

and indicate the definition of the differential operator

$$\mathbf{v} \cdot \nabla g = \sum_{j=1}^{d} \mathbf{v}_j \frac{\partial g}{\partial x_j}.$$

In the rest of this article, we denote vectors by bold roman letters and we use bold greek letters to denote second order tensors.

5.2.2 Mass balance equations

We assume that each component has its own velocity and the component 1 is proliferating with rate G that depends on the local pressure inside the mixture $p(\rho, c)$. The constitutive relation for this pressure term will be given in the following sections, however we just specify that p is not an unknown but it is a given function of ρ and c. Therefore, we have the mass balance equations for each component

$$\begin{cases} \frac{\partial \phi_1}{\partial t} + \operatorname{div}(\phi_1 \mathbf{v}_1) = \phi_1 G(p), \\ \frac{\partial \phi_2}{\partial t} + \operatorname{div}(\phi_2 \mathbf{v}_2) = 0. \end{cases} \tag{5.5}$$

Summing the two equations, we obtain the continuity equation for the total density of the mixture and using the mass fractions (denoting $c_1 = c$) and the relations (5.2), we obtain the balance equation for the density of the mixture

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{v}) = \rho c G(p). \tag{5.6}$$

To obtain a system analogous to (5.5), we rewrite the first equation of (5.5) using the definition of the mass fraction (5.2) to obtain

$$\frac{\partial \rho c}{\partial t} + \operatorname{div}(\rho c \mathbf{v}_1) = \rho c G(p). \tag{5.7}$$

The mass of the component 1 is transported by the average velocity $\mathbf{v}$ and the remaining diffusive flux $\mathbf{J}_1 = \rho c (\mathbf{v} - \mathbf{v}_1)$. Therefore, we can replace the previous equation by

$$\frac{\partial \rho c}{\partial t} + \operatorname{div}(\rho c \mathbf{v}) = \operatorname{div}(\mathbf{J}_1) + \rho c G(p).$$

Then, using the definition of the material derivative (5.4) and the mass balance equation for the total mixture (5.6), the left-hand side of the previous equation reads

$$\frac{\partial \rho c}{\partial t} + \operatorname{div}(\rho c \mathbf{v}) = \rho \frac{\mathrm{D}c}{\mathrm{D}t} + c \left[\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{v}) \right] = \rho \frac{\mathrm{D}c}{\mathrm{D}t} + \rho c^2 G(p).$$

Altogether, we obtain the balance equation for the mass fraction of the component 1

$$\rho \frac{\mathrm{D}c}{\mathrm{D}t} = \operatorname{div}(\mathbf{J}_1) + \rho c (1 - c) G(p). \tag{5.8}$$

Since $c_2 = 1 - c$, solving the equations (5.6) and (5.8) is equivalent to solving the system (5.5). In the following, we refer to c as the order parameter (terminology often use in the framework of the Cahn-hilliard model [48, 47]).

5.2.3 Balance of linear momentum

We write the balance of linear momentum [73], which describes the evolution of the velocity $\mathbf{v}$ due to internal stresses. Indeed, we neglect the effect of any external forces, including gravity. Following continuum mechanics, the Cauchy stress tensor gives the stresses acting inside the mixture due to viscous and non-viscous effects. An additional stress must be taken into account to represent the effect of concentration gradients [80]. Altogether, we assume that the stress tensor is a function of the total density ρ, the order parameter c (i.e. the mass fraction of population 1), its gradient ∇c, and the total velocity of the mixture $\mathbf{v}$ i.e.

$$\boldsymbol{\sigma} = \boldsymbol{\sigma}(\rho, c, \nabla c, \mathbf{v}).$$

The friction around the pores is modeled by a drag term in the balance equation [151] with a friction coefficient $\kappa(c) = \nu(c)/C_p(c)$, where $\nu(c) = \nu_1 c + \nu_2(1-c)$ is the weighted viscosity (with ν_1 and ν_2 being the viscosities of fluid 1 and 2 respectively) and $C_p(c)$ is the weighted conductance of the pores. The conductance is given by

$$C_p(c) = \frac{\xi(c)\rho}{\nu(c)}, \tag{5.9}$$

where $\xi(c)$ is the weighted permeability of the porous medium.

For each dimension (for example if $d = 3$, then $j = \{x, y, z\}$), the balance of linear momentum reads [73]

$$\frac{\partial \rho \mathbf{v}_j}{\partial t} + \operatorname{div}\left(\rho \mathbf{v}_j \mathbf{v}\right) = \operatorname{div}\left(\boldsymbol{\sigma}\right)_j - \kappa(c)\mathbf{v}_j.$$

Then, using the continuity equation (5.6), we can rearrange the left-hand side to obtain

$$\frac{\partial \rho \mathbf{v}_j}{\partial t} + \operatorname{div}\left(\rho \mathbf{v}_j \mathbf{v}\right) = \rho \frac{\mathrm{D}\mathbf{v}_j}{\mathrm{D}t} + \mathbf{v}_j \left[\frac{\partial \rho}{\partial t} + \operatorname{div}\left(\rho \mathbf{v}\right)\right] = \rho \frac{\mathrm{D}\mathbf{v}_j}{\mathrm{D}t} + \rho c \mathbf{v}_j G(p).$$

Therefore, we have

$$\rho \frac{\mathrm{D}\mathbf{v}_j}{\mathrm{D}t} = \operatorname{div}(\boldsymbol{\sigma})_j - \kappa(c)\mathbf{v}_j - \rho c \mathbf{v}_j G(p).$$

Then, we can rewrite the balance of linear momentum for each direction in a more compact form

$$\rho \frac{\mathrm{D}\mathbf{v}}{\mathrm{D}t} = \operatorname{div}(\boldsymbol{\sigma}) - \kappa(c)\mathbf{v} - \rho c \mathbf{v} G(p). \tag{5.10}$$

5.2.4 Energy balance

The total energy of the mixture is the sum of the kinetic energy $\rho \frac{1}{2} |\mathbf{v}|^2$ and of the internal energy ρu, where $u = u(\rho, c, \nabla c)$ is a specific internal energy. Comparing to the classical conservation law for the total energy, we have an additional energy flux $\boldsymbol{\tau} \frac{\mathrm{D}c}{\mathrm{D}t}$. Indeed, due to the interfacial region, surface effects must be taken into account. Following this direction, Gurtin [109] proposed to include in the second law of thermodynamics, the effect of an additional force called the *microscopic-stress* and is related to forces acting at the microscopic scale. We denote this supplementary stress by $\boldsymbol{\tau}$.

Since we assume that the system is maintained in an isothermal state, the balance equation

for the energy is given by [73]

$$\frac{\partial}{\partial t}\left(\rho\frac{1}{2}\left|\mathbf{v}\right|^2+\rho u\right)+\operatorname{div}\left(\rho\left(\frac{1}{2}\left|\mathbf{v}\right|^2+u\right)\mathbf{v}\right)=\operatorname{div}\left(\boldsymbol{\sigma}^T\mathbf{v}\right)+\operatorname{div}\left(\boldsymbol{\tau}\frac{\mathrm{D}c}{\mathrm{D}t}\right)-\operatorname{div}\left(\mathbf{q}\right)+\rho g, \quad (5.11)$$

where $\mathbf{q}$ is the heat flux and ρg is the density of heat sources to maintain the temperature constant. Then, repeating the same calculations on the left-hand side to use the balance of mass (5.6), we get

$$\frac{\partial}{\partial t}\left(\rho\frac{1}{2}\left|\mathbf{v}\right|^2+\rho u\right)+\operatorname{div}\left(\rho\left(\frac{1}{2}\left|\mathbf{v}\right|^2+u\right)\mathbf{v}\right)=\rho\left[\frac{\mathrm{D}}{\mathrm{D}t}\left(\frac{1}{2}\left|\mathbf{v}\right|^2+u\right)\right]+\left(\frac{1}{2}\left|\mathbf{v}\right|^2+u\right)\rho cG(p)$$

Applying the chain rule to the kinetic part, we get

$$\rho\frac{\mathrm{D}}{\mathrm{D}t}\left(\frac{1}{2}\left|\mathbf{v}\right|^2\right)=\rho\mathbf{v}\cdot\frac{\mathrm{D}\mathbf{v}}{\mathrm{D}t},$$

and using the balance of linear momentum (5.10), we obtain

$$\rho\mathbf{v}\cdot\frac{\mathrm{D}\mathbf{v}}{\mathrm{D}t}=\mathbf{v}\cdot\operatorname{div}(\boldsymbol{\sigma})-\kappa(c)\left|\mathbf{v}\right|^2-\rho c\left|\mathbf{v}\right|^2G(p)$$

Using this previous results inside (5.11), we obtain the balance equation for the internal energy

$$\rho\frac{\mathrm{D}u}{\mathrm{D}t}=\operatorname{div}\left(\boldsymbol{\sigma}^T\mathbf{v}\right)-\mathbf{v}\cdot\operatorname{div}\left(\boldsymbol{\sigma}\right)+\operatorname{div}\left(\boldsymbol{\tau}\frac{\mathrm{D}c}{\mathrm{D}t}\right)+\kappa(c)\left|\mathbf{v}\right|^2-\operatorname{div}\left(\mathbf{q}\right)+\rho g+\rho c\left(\frac{1}{2}\left|\mathbf{v}\right|^2-u\right)G(p).$$

However, since

$$\mathbf{v}\cdot\left(\operatorname{div}\left(\boldsymbol{\sigma}\right)\right)-\operatorname{div}\left(\boldsymbol{\sigma}^T\mathbf{v}\right)=-\boldsymbol{\sigma}:\nabla\mathbf{v},$$

where $\nabla\mathbf{v}=\left(\partial_{x_j}\mathbf{v}_i\right)_{i,j=1,\dots,d}$ is the Jacobi matrix and, we have $A:B=\sum_{i,j}A_{ij}B_{ij}$, for two matrices A,B. Altogether, we have the balance equation for the internal energy

$$\rho\frac{\mathrm{D}u}{\mathrm{D}t}=\boldsymbol{\sigma}:\nabla\mathbf{v}+\operatorname{div}\left(\boldsymbol{\tau}\frac{\mathrm{D}c}{\mathrm{D}t}\right)+\kappa(c)\left|\mathbf{v}\right|^2-\operatorname{div}\left(\mathbf{q}\right)+\rho g+\rho c\left(\frac{1}{2}\left|\mathbf{v}\right|^2-u\right)G(p). \quad (5.12)$$

5.2.5 Entropy balance and Clausius-Duhem inequality

We aim at applying the second law of thermodynamics. To do so, we define the entropy $s=s(\rho,c,\nabla c)$ and the Helmholtz free energy $\mathcal{F}=\mathcal{F}(\rho,c,\nabla c)$, both related through the equation

$$\mathcal{F}=u-Ts, \quad (5.13)$$

where T denotes the temperature.

From the mass balance equation (5.6), we have the entropy balance equation

$$\frac{\partial\rho s}{\partial t}+\operatorname{div}(s\rho\mathbf{v})=\rho\frac{\mathrm{D}s}{\mathrm{D}t}+s\left[\frac{\partial\rho}{\partial t}+\operatorname{div}\left(\rho\mathbf{v}\right)\right]=\rho\frac{\mathrm{D}s}{\mathrm{D}t}+\rho csG(p). \quad (5.14)$$

Then, using the definition of the Helmholtz free energy (5.13) and the balance of energy (5.12),

we obtain

$$\begin{aligned}\rho\frac{\mathrm{D}s}{\mathrm{D}t} &= -\frac{\rho}{T}\frac{\mathrm{D}\mathcal{F}}{\mathrm{D}t} + \frac{\rho}{T}\frac{\mathrm{D}u}{\mathrm{D}t} \\ &= -\frac{\rho}{T}\frac{\mathrm{D}\mathcal{F}}{\mathrm{D}t} + \frac{1}{T}\left[\boldsymbol{\sigma} : \nabla\mathbf{v} + \mathrm{div}\left(\boldsymbol{\tau}\frac{\mathrm{D}c}{\mathrm{D}t}\right) + \kappa(c)\left|\mathbf{v}\right|^2 - \mathrm{div}\left(\mathbf{q}\right) + \rho g + \rho c\left(\frac{1}{2}\left|\mathbf{v}\right|^2 - u\right)G(p)\right],\end{aligned} \tag{5.15}$$

where we have replaced the material derivative of the internal energy using its balance equation (5.12).

The constitutive relations for the functions constituting the Navier-Stokes-Cahn-Hilliard model are often derived to satisfy the Clausius-Duhem inequality (Coleman-Noll Procedure) [73]. Indeed, this inequality provides a set of restrictions for the dissipative mechanisms occurring in the system. However, in our case, due to the presence of source terms, we can not ensure that this inequality holds without some assumptions on the proliferation and the friction of the fluid around the pores. Therefore, we use here a different method: the Lagrange multipliers method. Indeed, Liu [137] and Müller [148] method is based on using Lagrange multipliers to derive a set of restrictions on the constitutive relations that can be applied even in the presence of source terms.

Following classical Thermodynamics [148], we state the second law as an entropy inequality, i.e., the Clausius-Duhem inequality in local form [73]

$$\rho\frac{\mathrm{D}s}{\mathrm{D}t} \geq -\mathrm{div}\left(\frac{\vec{q}\cdot\mathbf{v}}{T}\right) + \frac{\rho g}{T} + \mathrm{div}\left(\mathcal{J}\right) + L_\rho\rho c G(p), \tag{5.16}$$

where $\mathcal{J}$ is the entropy flux and L_ρ is an unknown Lagrange multiplier associated to the total mass increase [110]. The inequality (5.16) results from the fact that the entropy of the mixture can only increase. Using the equation (5.15), we obtain

$$\frac{\rho}{T}\frac{D\mathcal{F}}{Dt} - \frac{1}{T}\left[\boldsymbol{\sigma} : \nabla\mathbf{v} + \mathrm{div}\left(\boldsymbol{\tau}\frac{\mathrm{D}c}{\mathrm{D}t}\right) + \kappa(c)\left|\mathbf{v}\right|^2 + \rho c\left(\frac{1}{2}\left|\mathbf{v}\right|^2 - u\right)G(p)\right] + \mathrm{div}\left(\mathcal{J}\right) + L_\rho\rho c G(p) \leq 0. \tag{5.17}$$

Then, using the chain rule

$$\frac{D\mathcal{F}}{Dt} = \frac{D\rho}{Dt}\frac{\partial\mathcal{F}}{\partial\rho} + \frac{Dc}{Dt}\frac{\partial\mathcal{F}}{\partial c} + \frac{D\nabla c}{Dt}\frac{\partial\mathcal{F}}{\partial\nabla c},$$

and

$$\frac{D\nabla c}{Dt} = \nabla\left[\frac{Dc}{Dt}\right] - \left(\nabla\mathbf{v}\right)^T\cdot\nabla c, \quad \frac{\mathrm{D}\rho}{\mathrm{D}t} = -\rho\mathrm{div}(\mathbf{v}) + \rho c G(p),$$

in the entropy inequality (5.17), we obtain

$$\begin{aligned}&\rho\left[\left(-\rho\mathrm{div}(\mathbf{v}) + \rho c G(p)\right)\frac{\partial\mathcal{F}}{\partial\rho} + \frac{\mathrm{D}c}{\mathrm{D}t}\frac{\partial\mathcal{F}}{\partial c} + \left(\nabla\left[\frac{Dc}{Dt}\right] - \left(\nabla\mathbf{v}\right)^T\cdot\nabla c\right)\frac{\partial\mathcal{F}}{\partial\nabla c}\right] \\ &- \left[\boldsymbol{\sigma} : \nabla\mathbf{v} + \mathrm{div}\left(\boldsymbol{\tau}\frac{\mathrm{D}c}{\mathrm{D}t}\right) + \kappa(c)\left|\mathbf{v}\right|^2 + \rho c\left(\frac{1}{2}\left|\mathbf{v}\right|^2 - u\right)G(p)\right] + T\mathrm{div}\left(\mathcal{J}\right) + L_\rho\rho c G(p) \leq 0.\end{aligned} \tag{5.18}$$

By chain rule, we have

$$\mathrm{div}\left(\boldsymbol{\tau}\frac{\mathrm{D}c}{\mathrm{D}t}\right) = \boldsymbol{\tau}\nabla\left[\frac{Dc}{Dt}\right] + \frac{\mathrm{D}c}{\mathrm{D}t}\mathrm{div}\left(\boldsymbol{\tau}\right).$$

Furthermore, we know that

$$-\rho^2 \mathrm{div}\,(\mathbf{v})\,\frac{\partial \mathcal{F}}{\partial \rho} = -\rho^2 \frac{\partial \mathcal{F}}{\partial \rho}\mathbf{1} : \nabla \mathbf{v},$$

and

$$-\rho\left((\nabla \mathbf{v})^T \cdot \nabla c\right)\frac{\partial \mathcal{F}}{\partial \nabla c} = -\rho\left(\nabla c \otimes \frac{\partial \mathcal{F}}{\partial \nabla c}\right) : \nabla \mathbf{v}.$$

Gathering the previous three relations and reorganizing the terms of (5.18), we obtain

$$\begin{aligned}\left(-\rho^2\frac{\partial \mathcal{F}}{\partial \rho}\mathbf{1} - \rho\nabla c \otimes \frac{\partial \mathcal{F}}{\partial \nabla c} - \boldsymbol{\sigma}\right) : \nabla \mathbf{v} + \left(\rho\frac{\partial \mathcal{F}}{\partial c} - \mathrm{div}(\boldsymbol{\tau})\right)\frac{\mathrm{D}c}{\mathrm{D}t} + \left(\rho\frac{\partial \mathcal{F}}{\partial \nabla c} - \boldsymbol{\tau}\right)\nabla\left[\frac{\mathrm{D}c}{\mathrm{D}t}\right] \\ + T\mathrm{div}\,(\mathcal{J}) + \rho c\left[L_\rho - \left(\frac{1}{2} + \kappa(c)\right)|\mathbf{v}|^2 + u + \rho\frac{\partial \mathcal{F}}{\partial \rho}\right] G(p) \leq 0.\end{aligned} \tag{5.19}$$

Then, we use Liu's Lagrange multipliers method [137]. We denote by L_c the Lagrange multiplier associated with the mass fraction equation (5.8). The method of Lagrange multipliers consists in setting the following local dissipation inequality that has to hold for arbitrary values of $(\rho, c, \nabla\rho, \nabla c, \mathbf{v}, p)$

$$\begin{aligned}-D_{\mathrm{iss}} := &\left(-\rho^2\frac{\partial \mathcal{F}}{\partial \rho}\mathbf{1} - \rho\nabla c \otimes \frac{\partial \mathcal{F}}{\partial \nabla c} - \boldsymbol{\sigma}\right) : \nabla \mathbf{v} \\ &+ \left(\rho\frac{\partial \mathcal{F}}{\partial c} - \mathrm{div}(\boldsymbol{\tau})\right)\frac{\mathrm{D}c}{\mathrm{D}t} + \left(\rho\frac{\partial \mathcal{F}}{\partial \nabla c} - \boldsymbol{\tau}\right)\nabla\left[\frac{\mathrm{D}c}{\mathrm{D}t}\right] + T\mathrm{div}\,(\mathcal{J}) \\ &+ \rho c\left[c_\rho - \left(\frac{1}{2} + \kappa(c)\right)|\mathbf{v}|^2 + u + \rho\frac{\partial \mathcal{F}}{\partial \rho}\right] G(p) \\ &- L_c\left(\rho\frac{\mathrm{D}c}{\mathrm{D}t} - \mathrm{div}\,(\mathbf{J}_1) - \rho c(1-c)G(p)\right) \leq 0.\end{aligned} \tag{5.20}$$

Since,

$$\mathrm{div}\,(L_c\mathbf{J}_1) = L_c\mathrm{div}\,(\mathbf{J}_1) + \nabla L_c \cdot \mathbf{J}_1,$$

we reorganize the terms of (5.20) to obtain

$$\begin{aligned}-D_{\mathrm{iss}} := &\left(-\rho^2\frac{\partial \mathcal{F}}{\partial \rho}\mathbf{1} - \rho\nabla c \otimes \frac{\partial \mathcal{F}}{\partial \nabla c} - \boldsymbol{\sigma}\right) : \nabla \mathbf{v} \\ &+ \left(\rho\frac{\partial \mathcal{F}}{\partial c} - \mathrm{div}(\boldsymbol{\tau}) - \rho L_c\right)\frac{\mathrm{D}c}{\mathrm{D}t} + \left(\rho\frac{\partial \mathcal{F}}{\partial \nabla c} - \boldsymbol{\tau}\right)\nabla\left[\frac{\mathrm{D}c}{\mathrm{D}t}\right] + \mathrm{div}\,(T\mathcal{J} + L_c\mathbf{J}_1) \\ &+ \rho c\left[L_\rho - \left(\frac{1}{2} + \kappa(c)\right)|\mathbf{v}|^2 + u + \rho\frac{\partial \mathcal{F}}{\partial \rho} - L_c(1-c)\right] G(p) \\ &- \nabla L_c \cdot \mathbf{J}_1 \leq 0.\end{aligned} \tag{5.21}$$

5.2.6 Constitutive assumptions and model equations

First of all, we assume that the free energy density $\mathcal{F}$ is of Ginzburg-Landau type has the following form [48, 47]

$$\mathcal{F}(\rho, c, \nabla c) := \psi_0(\rho, c) + \frac{\gamma}{2}|\nabla c|^2, \tag{5.22}$$

where ψ_0 is the homogeneous free energy accounting for the processes of phase separation and the gradient term $\frac{\gamma}{2}|\nabla c|^2$ represents the surface tension between the two phases. This free energy is the basis of the Cahn-Hilliard model which describes the phase separation occurring in binary mixtures (see [147] for a complete review on the Cahn-Hilliard model and its variants).

To satisfy the inequality (5.21), we first choose

$$\boldsymbol{\tau} \coloneqq \rho \frac{\partial \mathcal{F}}{\partial \nabla c} = \gamma \rho \nabla c.$$

Then, we define the chemical potential $\mu(\rho, c, \nabla c)$ by

$$\mu \coloneqq \frac{\partial \mathcal{F}}{\partial c} - \frac{1}{\rho}\mathrm{div}(\boldsymbol{\tau}) = \frac{\partial \mathcal{F}}{\partial c} - \frac{1}{\rho}\mathrm{div}(\rho \frac{\partial \mathcal{F}}{\partial \nabla c}) = \frac{\partial \psi_0}{\partial c} - \frac{\gamma}{\rho}\mathrm{div}\left(\rho \nabla c\right),$$

which in turn gives a condition for the Lagrange multiplier

$$L_c = \mu. \tag{5.23}$$

Using these previous constitutive relations, we have already canceled some terms in the entropy inequality

$$\left(\rho \frac{\partial \mathcal{F}}{\partial c} - \mathrm{div}(\boldsymbol{\tau}) - \rho L_c\right) \frac{\mathrm{D}c}{\mathrm{D}t} + \left(\rho \frac{\partial \mathcal{F}}{\partial \nabla c} - \boldsymbol{\tau}\right) \nabla \left[\frac{\mathrm{D}c}{\mathrm{D}t}\right] = 0.$$

Then, using classical results on isothermal diffusion [141, 73], we have

$$\mathcal{J} \coloneqq -\frac{\mu \mathbf{J}_1}{T}, \tag{5.24}$$

and using a generalized Fick's law, we have

$$\mathbf{J}_1 \coloneqq b(c)\nabla \mu, \tag{5.25}$$

where $b(c)$ is a nonnegative motility function that we will specify in the following. The two constitutive relations for the diffusive fluxes (5.24) and (5.25) together with (5.23), we obtain

$$\mathrm{div}\left(T\mathcal{J} + L_c \mathbf{J}_1\right) - \nabla L_c \cdot \mathbf{J}_1 = -b(c)\left|\nabla \mu\right|^2 \leq 0.$$

Following [141, 1], we define the pressure inside the mixture by

$$p \coloneqq \rho^2 \frac{\partial \psi_0}{\partial \rho}. \tag{5.26}$$

From standard rheology, we assume that the fluid meet Newton's rheological laws. The stress tensor is composed of two parts for the viscous $\tilde{\mathbf{P}}$ and non-viscous $\mathbf{P}$ contributions of stress

$$\boldsymbol{\sigma} \coloneqq \mathbf{P} + \tilde{\mathbf{P}}, \tag{5.27}$$

and we have by standard continuum mechanics [73]

$$\begin{cases} \mathbf{P} = -p\mathbf{1} - \gamma \rho \nabla c \otimes \nabla c, \\ \tilde{\mathbf{P}} = \nu(c)\left(\nabla \mathbf{v} + \nabla \mathbf{v}^T\right) + \lambda(c)\left(\mathrm{div}\left(\mathbf{v}\right)\right)\mathbf{1}, \end{cases} \tag{5.28}$$

where $\nu(c) = c\nu_1 + (1-c)\nu_2$ is the weighted viscosity coefficient and $\lambda(c) = c\lambda_1 + (1-c)\lambda_2$

is the weighted bulk viscosity with $\nu(c)$ positive and $|\lambda(c)| \leq \nu(c)$. The second term in the non-viscous part of the stress (namely $-\gamma(\rho\nabla c \otimes \nabla c)$) is representing capillary stresses that act at the interface of the two populations. Furthermore, we assume that the bulk viscosity is zero and, consequently, we set $\lambda(c) = -\frac{2}{3}\nu(c)$. This form for the stress tensor is also the same used for Navier-Stokes fluids [80].

Using (5.28), we can cancel a new term in (5.21)

$$\left(-\rho^2\frac{\partial \mathcal{F}}{\partial \rho}\mathbf{1} - \rho\nabla c \otimes \frac{\partial \mathcal{F}}{\partial \nabla c} - \boldsymbol{\sigma}\right) : \nabla \mathbf{v} = 0.$$

Therefore, the remaining terms of the entropy inequality are the ones associated with proliferation and friction. The last step to satisfy the entropy inequality is to choose arbitrarily a value for the Lagrange multiplier c_ρ, such that

$$\rho c\left[L_\rho - \left(\frac{1}{2} + \kappa(c)\right)|\mathbf{v}|^2 + u + \rho\frac{\partial \mathcal{F}}{\partial \rho} - \lambda_c(1-c)\right] G(p) \leq 0.$$

The obvious choice is of course

$$L_\rho = \left(\frac{1}{2} + \kappa(c)\right)|\mathbf{v}|^2 - u - \rho\frac{\partial \mathcal{F}}{\partial \rho} + \lambda_c(1-c).$$

From the previous constitutive relations and choices of Lagrange multipliers, we have that the dissipation inequality (5.21) is satisfied.

5.2.7 Summary of the model equations

Using the previous constitutive relations our model is the following generalized Navier-Stokes-Cahn-Hilliard (G-NSCH) system

$$\begin{aligned}
\frac{\partial \rho}{\partial t} &= -\mathrm{div}\,(\rho\mathbf{v})\,\rho c G(p),\\
\rho\frac{\mathrm{D}c}{\mathrm{D}t} &= \mathrm{div}\,(b(c)\nabla\mu) + \rho c(1-c)G(p),\\
\rho\mu &= -\gamma\mathrm{div}\,(\rho\nabla c) + \rho\frac{\partial \psi_0}{\partial c},\\
\rho\frac{\mathrm{D}\mathbf{v}}{\mathrm{D}t} &= -\left[\nabla p + \gamma\mathrm{div}\,(\rho\nabla c \otimes \nabla c)\right] + \mathrm{div}\left(\nu(c)\left(\nabla\mathbf{v} + \nabla\mathbf{v}^T\right)\right)\\
&\qquad - \frac{2}{3}\nabla\left(\nu(c)\left(\mathrm{div}\,(\mathbf{v})\right)\right) - \kappa(c)\mathbf{v} - \rho c\mathbf{v}G(p),
\end{aligned} \tag{5.29}$$

5.3 General assumptions and biologically relevant choice of the model functions

5.3.1 General forms and assumptions

The motility is a positive function of the order parameter (mass fraction) c. Hence, we assume that

$$b \in C^1([0,1];\mathbb{R}^+), \quad \text{and} \quad b(c) > 0 \quad \text{for} \quad 0 \leq c \leq 1. \tag{5.30}$$

In agreement with the literature (see e.g [58]) the homogeneous free energy $\psi_0(\rho, c)$ is assumed

to be of the form

$$\psi_0(\rho, c) = \psi_e(\rho) + H(c)\log\rho + Q(c), \tag{5.31}$$

where $H(\cdot)$ and $Q(\cdot)$ satisfy

$$\begin{aligned} -H_1 \leq H'(c), \quad H(c) \leq H_2, \quad c \in \mathbb{R}, \quad H_1, H_2 > 0, \\ Q_1\,|c| - Q_2 \leq Q'(c) \leq Q_3(1+|c|), \quad c \in \mathbb{R}, \quad Q_1, Q_2, Q_3 > 0. \end{aligned} \tag{5.32}$$

Then, using the constitutive relation for the pressure we have

$$p(\rho, c) = \rho^2 \frac{\partial \psi_0}{\partial \rho} = p_e(\rho) + \rho H(c), \tag{5.33}$$

where $p_e = \rho^2 \psi_e'(\rho)$ and is assumed to satisfy

$$p_1 \rho^{a-1} - p_2 \leq p_e'(\rho) \leq p_3(1+\rho^{a-1}), \quad \text{for} \quad a > 3/2, \quad p_1, p_2, p_3 > 0. \tag{5.34}$$

The growth function $G(p)$ is used to represent the capacity of cells to divide accordingly to the pressure exerted on them. It is well known that cells are able to divide as long as the pressure is not too large. Once a certain pressure $p_{\max}$ is reached cells enter a quiescent state. Therefore, we assume that

$$G'(p) \leq 0, \quad \text{and} \quad G(p) = 0 \quad \text{for} \quad p > p_{\max}. \tag{5.35}$$

5.3.2 Biologically consistent choice of functions

As said in the derivation of the model, the free energy density $\mathcal{F}$ is the sum of two terms: $\frac{\gamma}{2}|\nabla c|^2$ taking into account the surface tension effects existing between the phases of the mixture and the potential $\psi_0(\rho, c)$ representing the cell-cell interactions and pressure. The function $b(\rho)$ is the active motility of the cells of the growing population.

Let us explain how the choices of functions for the free energy density and mobility are motivated by biological observations.

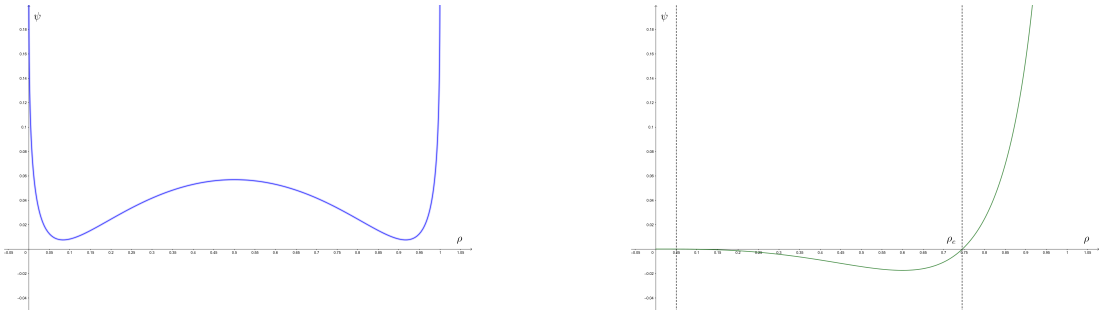
To satisfy the conditions (5.30), we propose to choose

$$b(\rho) = C_b c(1-c)^\alpha, \quad \alpha \geq 1, \tag{5.36}$$

where C_b is a positive constant. We use for the pressure a power law such that

$$p_e(\rho) = \frac{1}{a-1}\rho^{a-1}. \tag{5.37}$$

For $H(c)$ and $G(c)$ two cases can be considered depending on the behavior of the cells we want to represent. If the two cell populations exert attractive forces when they recognize cells of the same type and repulsion with the other type, this potential has to take a form of a double-well where the two stable phases are located at the bottom of the two wells (see e.g. Figure 5.2a). This is a situation close to the phase separation in binary fluids. Thermodynamically consistent potentials are of Ginzburg-Landau type with presence logarithmic terms. Even though the double-well form of the potential is originally used for applications dealing with materials, it can also be motivated for biological purposes. Indeed, considering an application where the mixture is saturated with two cell types and are of similar compressibility, a double-well potential is biologically relevant and reflects correctly the expected behavior of cells: they are attracted to each other respectively to their cell type at low densities and after a certain density they start to repel each others to avoid the creation of overcrowded zones. A typical example of biologically



(a) (b)

Figure 5.2 – Double-well logarithmic potential (left) and single-well logarithmic potential (right)

relevant double-well potential is given by

$$H(c) = \frac{1}{2}\left((1-c)\log(1-c) + c\log(c)\right), \quad G(c) = -\frac{\theta}{2}(c-\frac{1}{2})^2 + k, \tag{5.38}$$

where $\theta > 1$ and k is an arbitrary constant.

To meet the phenomenological observations of interaction between cells when the mixture is composed of only one cell population (and the other component of the mixture is supposed to be much more compressible), a single-well potential seems more appropriate [46, 56].

Indeed, when the distance between cells falls below a certain value (i.e. if the cell density is large enough), cells are attracted to each other. Then, it exists a threshold value called the mechanical equilibrium for which $\rho H(c_e) + C(c_e) = 0$ i.e. there is an equilibrium between attractive and repulsive forces. For larger cell densities, cells are packed too close to each others, they thus experience a repulsive force. when cells are so packed that they fill the whole control volume, then the repulsive force becomes infinite due to the pressure. The representation of such functional is depicted in Figure 5.2b. A typical example of single-well potential which has been use for the modelling of living tissue and cancer [56, 8] is

$$H(c) = -(1-c_e)\log(1-c), \quad G(c) = -\frac{c^3}{3} - (1-c_e)\frac{c^2}{2} - (1-c_e)c + k, \tag{5.39}$$

where k is an arbitrary constant.

The growth function

$$G(p) = \frac{200}{\pi}\arctan\left(4\left[p_{\max} - p\right]_+\right),$$

where $p_{\max} = p(\rho_{\max}, c)$, meet the conditions (5.35).

5.3.3 Non-dimensionalized model

We introduce the dimensionless independent variables $\overline{x} = x/L^\star$, $\overline{t} = V^\star t/L^\star$, where $L^\star$ and $V^\star$ are the characteristic scale of length and velocity. Therefore, we rescale the dependent variables by $\overline{\mathbf{v}} = \mathbf{v}/V^\star$, $\overline{\rho} = \rho/\rho^\star$, $\overline{p} = p(\rho^\star\mu^\star)$ and $\overline{\psi}_0 = \psi_0/\mu^\star$, where $\rho^\star$ and $\mu^\star$ are characteristic

quantities. Dropping the bar notation, we obtain the non-dimensionalized model

$$\begin{cases} \frac{\partial \rho}{\partial t} = & -\mathrm{div}\,(\rho \mathbf{v}) + C_{LV}\rho c S_1(p), \\ \rho \frac{Dc}{Dt} = & \mathrm{div}\left(\frac{1}{Pe(c)}\nabla \mu\right) + C_{LV}\rho c(1-c)S_1(p), \\ \rho\mu = & -C_\varepsilon \mathrm{div}\,(\rho \nabla c) + \rho \frac{\partial \psi_0}{\partial c}, \\ \rho \frac{D\mathbf{v}}{Dt} = & -\frac{1}{M}\left[\nabla p + C_\varepsilon \mathrm{div}\,(\rho \nabla c \otimes \nabla c)\right] + \mathrm{div}\left(\frac{1}{Re(c)}\left(\nabla \mathbf{v} + \nabla \mathbf{v}^T\right)\right) \\ & -\frac{2}{3}\nabla\left(\frac{1}{Re(c)}\left(\nabla \cdot \mathbf{v}\right)\right) - C_{LV}\mathbf{v}\left(\kappa(c) + \rho c S_1(p)\right), \end{cases} \tag{5.40}$$

where we have used the definition of the concentration dependent of the Peclet and Reynold number $Pe(c) = \frac{\rho^* L^* V^*}{b(c)}$ and $Re(c) = \frac{\rho^* L^* V^*}{\nu(c)}$. The other constants are defined by $C_{LV} = \frac{L^*}{V^*}$, $C_\varepsilon = \frac{\varepsilon}{\mu^\star (L^*)^2}$, and $M = \frac{(V^*)^2}{\mu^\star}$. In the same manner, we non-dimensionalize the energy of the fluid

$$\mathcal{E}(t) = \int_\Omega \rho \left[\frac{1}{2}\mathbf{v}^2 + \psi_0(\rho, c) + \frac{\varepsilon}{2}\left|\nabla c\right|^2\right] \mathrm{d}x,$$

using the rescaling $\overline{\mathcal{E}} = \frac{\mathcal{E}}{(L^\star)^3 \rho^\star (V^\star)^2}$ to obtain

$$\mathcal{E}(t) = \int_\Omega \rho \left[\frac{1}{2}\mathbf{v}^2 + C_V \psi_0(\rho, c) + \frac{C_\varepsilon}{2}\left|\nabla c\right|^2\right] \mathrm{d}x,$$

where we have dropped the bar notation.

5.4 Large friction hypothesis

We aim at studying the asymptotic model recovered when we assume that the effects of friction around the pores and of the pressure inside the fluid are large compared to the other effects taken into account in the model (5.29). We also assume that the inertia effects of the total mixture are negligible leading to

$$\rho \frac{\mathrm{D}\mathbf{v}}{\mathrm{D}t} = 0.$$

To account for the fact that friction around and the pressure are the predominant effect, we use a small positive parameter $\varepsilon << 1$ and define the rescaled coefficients and functions

$$\kappa(c) = \frac{\kappa_\varepsilon(c_\varepsilon)}{\varepsilon}, \quad \nu(c) = \varepsilon^2 \nu_\varepsilon(c_\varepsilon), \quad G(p) = \varepsilon G_\varepsilon(p_\varepsilon), \quad b(c) = \varepsilon b_\varepsilon(c_\varepsilon). \tag{5.41}$$

Then, we rescale time accordingly to

$$t = \frac{t_\varepsilon}{\varepsilon},$$

which induces for every time dependent variable

$$\partial_t \rho = \varepsilon \partial_{t_\varepsilon} \rho_\varepsilon, \quad \partial_t c = \varepsilon \partial_{t_\varepsilon} c_\varepsilon, \quad \partial_t \mathbf{v} = \varepsilon \partial_{t_\varepsilon} \mathbf{v}_\varepsilon.$$

Using this rescaling and neglecting inertia effects, we obtain the following modified momentum equation

$$0 = -\left[\nabla p_\varepsilon + \varepsilon^2\gamma \mathrm{div}\left(\rho_\varepsilon \nabla c_\varepsilon \otimes \nabla c_\varepsilon\right)\right] + \varepsilon^2 \mathrm{div}\left(\nu_\varepsilon(c_\varepsilon)\left(\nabla \mathbf{v}_\varepsilon + \nabla \mathbf{v}_\varepsilon^T\right)\right) - \varepsilon^2 \frac{2}{3}\nabla\left(\nu_\varepsilon(c_\varepsilon)\left(\nabla \cdot \mathbf{v}_\varepsilon\right)\right) - \frac{\kappa_\varepsilon(c_\varepsilon)}{\varepsilon}\mathbf{v}_\varepsilon - \varepsilon\rho_\varepsilon c_\varepsilon \mathbf{v} G_\varepsilon(p_\varepsilon).$$

Neglecting the second order terms in ε, we obtain the definition of the velocity field

$$\mathbf{v}_\varepsilon = -\frac{\nabla p_\varepsilon}{\frac{\kappa_\varepsilon(c_\varepsilon)}{\varepsilon} + \varepsilon\rho_\varepsilon c_\varepsilon G_\varepsilon(p_\varepsilon)}. \tag{5.42}$$

This form of the velocity field bears some resemblance with the well-known Darcy's Law [66]. The rescaling is also applied to the other equations of system (5.29), so using the definition of the velocity field (5.42), we obtain for the total density

$$\varepsilon\frac{\partial \rho_\varepsilon}{\partial t_\varepsilon} - \varepsilon \mathrm{div}\left(\frac{\rho_\varepsilon}{\kappa_\varepsilon(c_\varepsilon) + \varepsilon^2\rho_\varepsilon c_\varepsilon G_\varepsilon(p_\varepsilon)}\nabla p_\varepsilon\right) = \varepsilon\rho_\varepsilon c_\varepsilon(1-c_\varepsilon)G_\varepsilon(p_\varepsilon),$$

and simplifying ε on both sides we obtain

$$\frac{\partial \rho_\varepsilon}{\partial t_\varepsilon} - \mathrm{div}\left(\frac{\rho_\varepsilon}{\kappa_\varepsilon(c_\varepsilon) + \varepsilon^2\rho_\varepsilon c_\varepsilon G_\varepsilon(p_\varepsilon)}\nabla p_\varepsilon\right) = \rho_\varepsilon c_\varepsilon(1-c_\varepsilon)G_\varepsilon(p_\varepsilon). \tag{5.43}$$

Then, for the mass fraction equation we have

$$\varepsilon\rho_\varepsilon\frac{\partial c_\varepsilon}{\partial t_\varepsilon} - \varepsilon\frac{\rho}{\kappa_\varepsilon(c_\varepsilon) + \varepsilon^2\rho_\varepsilon c_\varepsilon G_\varepsilon(p_\varepsilon)}\nabla p_\varepsilon \cdot \nabla c_\varepsilon = \varepsilon \mathrm{div}\left(b_\varepsilon(c_\varepsilon)\nabla\mu\right) + \varepsilon\rho_\varepsilon c_\varepsilon G_\varepsilon(p_\varepsilon),$$

and again simplifying ε on both sides we find

$$\rho_\varepsilon\frac{\partial c_\varepsilon}{\partial t_\varepsilon} - \frac{\rho}{\kappa_\varepsilon(c_\varepsilon) + \varepsilon^2\rho_\varepsilon c_\varepsilon G_\varepsilon(p_\varepsilon)}\nabla p_\varepsilon \cdot \nabla c_\varepsilon = \mathrm{div}\left(b_\varepsilon(c_\varepsilon)\nabla\mu\right) + \rho_\varepsilon c_\varepsilon G_\varepsilon(p_\varepsilon), \tag{5.44}$$

Lastly, letting $\varepsilon \to 0$ in (5.43) and (5.44), we obtain for the zero order terms of the expansions for the different terms the limit system

$$\begin{cases} \frac{\partial \rho}{\partial t} - \mathrm{div}\left(\frac{\rho}{\kappa(c)}\nabla p\right) = \rho c G(p), \\ \rho\frac{\partial c}{\partial t} - \frac{\rho}{\kappa(c)}\nabla p \cdot \nabla c = \mathrm{div}\left(b(c)\nabla\frac{\partial \psi_0}{\partial c}\right) + \rho c(1-c)G(p). \end{cases} \tag{5.45}$$

This model is interesting to study because it features two important effects that drive the movement of the two phases. The gradient of the pressure gives the direction of the diffusion movement while the two components of the mixture experience attraction and repulsion due to the term $\nabla\frac{\partial \psi_0}{\partial c}$.

However, we can also deduce another model using the definition (5.42) and (5.3)

$$\mathbf{v}_\varepsilon = c_\varepsilon \mathbf{v}_{1,\varepsilon} + (1-c_\varepsilon)\mathbf{v}_{2,\varepsilon} = -\frac{\nabla p_\varepsilon}{\frac{c_\varepsilon\kappa_1 + (1-c_\varepsilon)\kappa_2}{\varepsilon} + \varepsilon\rho_\varepsilon c_\varepsilon G_\varepsilon(p_\varepsilon)}.$$

Multiplying the previous equation by $\frac{\varepsilon(c_\varepsilon\kappa_1+(1-c_\varepsilon)\kappa_2)}{\varepsilon(c_\varepsilon\kappa_1+(1-c_\varepsilon)\kappa_2)}$, we have

$$-\frac{\varepsilon\left(c_\varepsilon\kappa_1+(1-c_\varepsilon)\kappa_2\right)}{\left(c_\varepsilon\kappa_1+(1-c_\varepsilon)\kappa_2\right)^2+\varepsilon^2\rho_\varepsilon c_\varepsilon G_\varepsilon(p_\varepsilon)\left(c_\varepsilon\kappa_1+(1-c_\varepsilon)\kappa_2\right)}\nabla p_\varepsilon.$$

Hence, we define

$$\begin{aligned}\mathbf{v}_{1,\varepsilon}&=-\frac{\varepsilon\kappa_1}{\left(c_\varepsilon\kappa_1+(1-c_\varepsilon)\kappa_2\right)^2+\varepsilon^2\rho_\varepsilon c_\varepsilon G_\varepsilon(p_\varepsilon)\left(c_\varepsilon\kappa_1+(1-c_\varepsilon)\kappa_2\right)}\nabla p_\varepsilon,\\ \mathbf{v}_{2,\varepsilon}&=-\frac{\varepsilon\kappa_2}{\left(c_\varepsilon\kappa_1+(1-c_\varepsilon)\kappa_2\right)^2+\varepsilon^2\rho_\varepsilon c_\varepsilon G_\varepsilon(p_\varepsilon)\left(c_\varepsilon\kappa_1+(1-c_\varepsilon)\kappa_2\right)}\nabla p_\varepsilon.\end{aligned}$$

Then, using these definition in the system (5.5) with the previous rescaling, we obtain we get the system

$$\begin{cases}\varepsilon\frac{\partial\phi_{1,\varepsilon}}{\partial t_\varepsilon}-\operatorname{div}\left(\frac{\varepsilon\kappa_1\phi_{1,\varepsilon}}{\kappa_\varepsilon(c_{1,\varepsilon})^2+\varepsilon^2\phi_{1,\varepsilon}G_\varepsilon(p_\varepsilon)\kappa_\varepsilon(c_{1,\varepsilon})}\nabla p_\varepsilon\right)=\varepsilon\phi_{1,\varepsilon}G_\varepsilon(p_\varepsilon),\\ \varepsilon\frac{\partial\rho_\varepsilon c_{2,\varepsilon}}{\partial t_\varepsilon}-\operatorname{div}\left(\frac{\varepsilon\kappa_2\phi_{2,\varepsilon}}{\kappa_\varepsilon(1-c_{2,\varepsilon})^2+\varepsilon^2\phi_{1,\varepsilon}G_\varepsilon(p_\varepsilon)\kappa_\varepsilon(1-c_{2,\varepsilon})}\nabla p_\varepsilon\right)=0.\end{cases}$$

Then, simplifying the factorized ε, letting $\varepsilon\to 0$, we obtain for the zero order terms the system

$$\begin{cases}\frac{\partial\phi_1}{\partial t}-\kappa_1\operatorname{div}\left(\frac{\phi_1}{\kappa(c_1)^2}\nabla p\right)=\phi_1G(p),\\ \frac{\partial\phi_2}{\partial t}-\kappa_2\operatorname{div}\left(\frac{\phi_2}{\kappa(c_2)^2}\nabla p\right)=0.\end{cases}\tag{5.46}$$

If we further assume that $\kappa(c_1)=\kappa_1$, and $\kappa(c_2)=\kappa_2$ (therefore assuming that the friction coefficient is a constant function), we get

$$\begin{cases}\frac{\partial\phi_1}{\partial t}-\frac{1}{\kappa_1}\operatorname{div}\left(\phi_1\nabla p\right)=\phi_1G(p),\\ \frac{\partial\phi_2}{\partial t}-\frac{1}{\kappa_2}\operatorname{div}\left(\phi_2\nabla p\right)=0.\end{cases}\tag{5.47}$$

As said in the introduction, the model (5.47) is the same as (5.1), and has been proposed previously in [139] to model the situation where a cell population is proliferating inside another non-proliferating one. The model also considers a contrast in mobility for the two populations. In [139], the authors defined κ_1^{-1} and κ_2^{-1} as mobility coefficient i.e., the quotient of permeability and viscosity. Indeed, coming back to the definition of the friction coefficient $\kappa(c)=\nu(c)/C_p(c)$ (and assuming again that $\kappa(c_i)=\kappa_i=\nu_i/C_{p,i}$ for $i=1,2$ are constants) and of the conductance of the pores (5.9), our two coefficients of mobility κ_1^{-1} and κ_2^{-1} are defined as

$$\kappa_1^{-1}=\frac{K_1}{\nu_1},\quad\text{and}\quad\kappa_2^{-1}=\frac{K_2}{\nu_2},$$

where K_1 and K_2 are the permeability coefficients for the two components. Numerical simulations of the model (5.1) suggests that when the mobility of the proliferating population is larger than that of the other one then the solution exhibits patterns similar to Saffman-Taylor instabilities. Therefore, our result indicates that the adhesion on the ECM and its capacity to move inside (its permeability) plays a crucial role in the emergence of irregularities at the surface of growing tumors.

5.5 Finite volume numerical scheme

Let us explain the details of the finite volume numerical scheme. We are using an upwind method to calculate the convective terms with a MUSCL reconstruction at the interface of the cells. Let us describe the numerical components of our scheme for the case $d = 2$.

The mesh. We use a structured grid of $N = N_i N_j$ cells (where N_i and N_j are the numbers of cells for each direction x and y). We denote the set of centroids by $J_C = \{(x_i, x_j) \in \Omega \quad | \quad i = 1, \ldots, N_i, \quad j = 1, \ldots, N_j\}$. The centroids are equally spaced by a parameter Δx in the x direction and Δy in the y direction.

Calculation of convective terms. To ensure the stability of our scheme and the positivity of both the fluid density ρ and the order parameter c, we use an upwind method to approximate the convective terms. A general example of calculation of a convective term is

$$(\mathbf{v} \cdot \nabla c)_{i,j} = \left(\mathbf{v}_x \frac{\partial c}{\partial x} + \mathbf{v}_y \frac{\partial c}{\partial y} \right)_{i,j}.$$

We approximate in each direction by

$$\left(\mathbf{v}_x \frac{\partial c}{\partial x} \right)_{i,j} = \max(0, \mathbf{v}_i)(\frac{\partial c}{\partial x})_{i+1/2,j} + \min(0, \mathbf{v}_i)(\frac{\partial c}{\partial x})_{i-1/2,j},$$

and we approximate the gradients at the interface by

$$(\frac{\partial c}{\partial x})_{i+1/2,j} = \frac{(\frac{\partial c}{\partial x})^L_{i+1/2,j} + (\frac{\partial c}{\partial x})^R_{i+1/2,j}}{2},$$

and

$$(\frac{\partial c}{\partial x})^L_{i+1/2,j} = \frac{c^L_{i+1,j} - c^L_{i-1,j}}{2\Delta x}, \quad (\frac{\partial c}{\partial x})^R_{i+1/2,j} = \frac{c^R_{i+1,j} - c^R_{i-1,j}}{2\Delta x}.$$

The other type of convective term that we need to deal with is of the form

$$\operatorname{div}(\rho \mathbf{v})_{i,j} = \frac{1}{V}_{i,j} \sum_{\sigma \in \Gamma_{i,j}} \tilde{F}_\sigma(\rho, \mathbf{v}) \vec{S}_\sigma,$$

where $\Gamma_{i,j}$ is the set of faces for the cell that has $x_{i,j}$ as centroid and $V_{i,j}$ denotes its volume. We also denoted the flux crossing the interface σ by F_σ and $\vec{S}_\sigma$ is the surface vector.

To describe how the flux is calculated by the upwind approach, let us take the example of the right interface of the cell i, j

$$\tilde{F}_{i+1/2,j}(\rho, \mathbf{v}) = \rho^L_{i+1/2,j} \max(0, \mathbf{v}_{i+1/2,j}) + \rho^R_{i+1/2,j} \min(0, \mathbf{v}_{i+1/2,j}).$$

MUSCL reconstruction. At each cell interface we need to approximate the density of the fluid ρ, its velocity $\mathbf{v}$ and the order parameter c. Since the solution is expected to display large gradients, we use a high order reconstruction at the interface. We use the MUSCL method that approximates the solution in each cell by a linear piecewise function. The main ingredients of this scheme is the use of a slope limiter to reconstruct the variables at both sides of each interface. If we denote by $u^L_{i+1/2,j}$ the variable located the right interface but inside the cell and $u^R_{i+1/2,j}$ the value at the right interface but outside the cell, we have

$$u^L_{i+1/2,j} = u_{i,j} + \frac{1}{2}\phi(r_{i,j})(u_{i+1,j} - u_{i,j}), \quad u^R_{i+1/2,j} = u_{i+1,j} - \frac{1}{2}\phi(r_{i+1,j})(u_{i+2,j} - u_{i+1,j}).$$

Discrete scheme for model (5.29). Using a first order approximation of the time derivative with a time step Δt and a positive number $k = 0, \ldots, N_T - 1$ such that $t^k = k\Delta t$, we have $\frac{\partial \rho}{\partial t} \approx \frac{\rho^{k+1} - \rho^k}{\Delta t}$. Thus, using the finite volume space discretization above, we obtain the fully discrete linear semi-implicit scheme for centroid $x_{i,j}$

$$\frac{\rho_{i,j}^{k+1} - \rho_{i,j}^k}{\Delta t} + \frac{1}{V_{i,j}} \sum_{\sigma \in \Gamma_i} \tilde{F}_\sigma(\rho^k, \mathbf{v}^k) S_\sigma = \rho^k c^k S_1(p^k),$$

$$\rho_{i,j}^{k+1} \left[\frac{c_{i,j}^{k+1} - c_{i,j}^k}{\Delta t} + \left(\mathbf{v}_x^k \frac{\partial c^k}{\partial x} + \mathbf{v}_y^k \frac{\partial c^k}{\partial y} \right)_{i,j} \right] = \frac{1}{V}_{i,j} \sum_{\sigma \in \Gamma_{i,j}} \left[\tilde{F}_\sigma(\rho^k, \nabla \mu^k) \vec{S}_\sigma \right] + \rho^k c^k (1 - c^k) S_1(p^k),$$

$$p_{i,j}^{k+1} = (\rho_{i,j}^{k+1})^\gamma + \rho_{i,j}^{k+1} H(c_{i,j}^{k+1}),$$

$$\rho_{i,j}^{k+1} \mu_{i,j}^{k+1} = -\varepsilon \sum_{\sigma \in \Gamma_{i,j}} \tilde{F}_\sigma(\rho, \nabla c^{k+1}) + \rho_{i,j}^{k+1} \left(\frac{\partial \psi_0(c)}{\partial c} \right)_{i,j}^{k+1},$$

$$\begin{aligned}
&\rho_{i,j}^{k+1} \left[\frac{\mathbf{v}_{i,j,x}^{k+1} - \mathbf{v}_{i,j,x}^k}{\Delta t} + \left(\mathbf{v}_x^k \frac{\partial \mathbf{v}_x^k}{\partial x} + \mathbf{v}_y^k \frac{\partial \mathbf{v}_x^k}{\partial y} \right)_{i,j} \right] \\
&\quad = -\frac{1}{V_{i,j}} \Big[\frac{1}{\Delta x} \left(p_{i+1/2,j}^{k+1} - p_{i-1/2,j}^{k+1} \right) + \varepsilon \rho_{i,j}^{k+1} \frac{1}{\Delta x} \left(\left(\frac{\partial c^{k+1}}{\partial x} \right)_{i+1/2,j}^2 - \left(\frac{\partial c^{k+1}}{\partial x} \right)_{i-1/2,j}^2 \right) \\
&\quad + \varepsilon \rho_{i,j}^{k+1} \frac{1}{\Delta y} \left(\left(\frac{\partial c}{\partial y} \frac{\partial c}{\partial x} \right)_{i,j+1/2}^{k+1} - \left(\frac{\partial c}{\partial y} \frac{\partial c}{\partial x} \right)_{i,j-1/2}^{k+1} \right) \Big] \\
&\quad + \frac{1}{\Delta x} \left(\tilde{F}_{i+1/2,j} \left(\frac{1}{Re(c^{k+1})}, \nabla \mathbf{v}_x^k \right) - F_{i-1/2,j} \left(\frac{1}{Re(c^{k+1})}, \nabla \mathbf{v}_x^k \right) \right) \\
&\quad + \frac{1}{\Delta y} \left(\tilde{F}_{i,j+1/2} \left(\frac{1}{Re(c^{k+1})}, \nabla \mathbf{v}_x^k \right) - F_{i,j-1/2} \left(\frac{1}{Re(c^{k+1})}, \nabla \mathbf{v}_x^k \right) \right) \\
&\quad - \frac{2}{3} \frac{1}{\Delta x} \left(\tilde{F}_{i+1/2,j} \left(\frac{1}{Re(c^{k+1})}, \nabla \cdot \mathbf{v}^k \right) - \tilde{F}_{i-1/2,j} \left(\frac{1}{Re(c^{k+1})}, \nabla \cdot \mathbf{v}^k \right) \right) \\
&\quad - \mathbf{v}_{i,j,x}^k \rho_{i,j}^{k+1} c_{i,j}^{k+1} S_1(p_{i,j}^{k+1}),
\end{aligned}$$

$$
\begin{aligned}
&\rho_{i,j}^{k+1}\left[\frac{\mathbf{v}_{i,j,y}^{k+1}-\mathbf{v}_{i,j,y}^{k}}{\Delta t}+\left(\mathbf{v}_x^k\frac{\partial \mathbf{v}_y^k}{\partial x}+\mathbf{v}_y^k\frac{\partial \mathbf{v}_y^k}{\partial y}\right)_{i,j}\right]\\
&=-\frac{1}{V_{i,j}}\Big[\frac{1}{\Delta y}\left(p_{i,j+1/2}^{k+1}-p_{i,j-1/2}^{k+1}\right)+\varepsilon\rho_{i,j}^{k+1}\frac{1}{\Delta y}\left(\left(\frac{\partial c^{k+1}}{\partial y}\right)^2_{i,j+1/2}-\left(\frac{\partial c^{k+1}}{\partial y}\right)^2_{i,j-1/2}\right)\\
&+\varepsilon\rho_{i,j}^{k+1}\frac{1}{\Delta x}\left(\left(\frac{\partial c}{\partial y}\frac{\partial c}{\partial x}\right)^{k+1}_{i+1/2,j}-\left(\frac{\partial c}{\partial y}\frac{\partial c}{\partial x}\right)^{k+1}_{i-1/2,j}\right)\Big]\\
&+\frac{1}{\Delta x}\left(\tilde{F}_{i+1/2,j}\left(\frac{1}{Re(c^{k+1})},\nabla\mathbf{v}_y^k\right)-F_{i-1/2,j}\left(\frac{1}{Re(c^{k+1})},\nabla\mathbf{v}_y^k\right)\right)\\
&+\frac{1}{\Delta y}\left(\tilde{F}_{i,j+1/2}\left(\frac{1}{Re(c^{k+1})},\nabla\mathbf{v}_y^k\right)-F_{i,j-1/2}\left(\frac{1}{Re(c^{k+1})},\nabla\mathbf{v}_y^k\right)\right)\\
&-\frac{2}{3}\frac{1}{\Delta y}\left(\tilde{F}_{i,j+1/2}\left(\frac{1}{Re(c^{k+1})},\nabla\cdot\mathbf{v}^k\right)-\tilde{F}_{i,j-1/2}\left(\frac{1}{Re(c^{k+1})},\nabla\cdot\mathbf{v}^k\right)\right)\\
&-\mathbf{v}_{i,j,y}^k\rho_{i,j}^{k+1}c_{i,j}^{k+1}S_1(p_{i,j}^{k+1}).
\end{aligned}
$$

Part III

Structure-preserving numerical method for nonlinear PDEs

Chapter 6

The Scalar Auxiliary Variable method for the volume-filling Keller-Segel model.

Abstract

We describe and analyze a finite element numerical scheme for the parabolic-parabolic Keller-Segel model. To prevent the blow-up of the cell density in finite time, we use the volume-filling modification of the chemosensitivity. The scalar auxiliary variable method is used to retrieve the monotonic decay of the energy associated with the system at the discrete level. This method relies on the interpretation of the Keller-Segel model as a gradient flow. The SAV finite-element scheme is stabilized by a simple upwind method. The resulting numerical scheme is efficient and easy to implement. We show the existence of a unique non-negative solution and that a modified discrete energy is obtained due to the use of the SAV method. From numerical simulations, we observe that the SAV-upwind scheme enhances the spatial accuracy compared to classical upwind methods.

This chapter contains preliminary results from an ongoing work.

6.1 Introduction

Since chemotaxis is observed very widely in various areas of biology and medicine, it becomes a prolific subject in mathematical biology throughout the past decades. Among the different mathematical models used to represent chemotaxis of living organisms, the Keller-Segel equation is one of the most recognized. It has been introduced by Keller and Segel [123] to depict the movement of the Dictyostelium discoideum toward the location of high concentration of adenosine 3', 5'-cyclic monophosphate. The parabolic-parabolic Keller-Segel model (KS in short) is often set in a bounded domain $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$ with a Lipschitz boundary $\partial\Omega$ and reads

$$\partial_t u = \nabla \cdot (D_u \nabla u - \chi_c \varphi(u) \nabla c) \quad \text{in} \quad \Omega \times (0, +\infty), \tag{6.1}$$

$$\tau \partial_t c = \Delta c - \alpha c + u \quad \text{in} \quad \Omega \times (0, +\infty), \tag{6.2}$$

endowed with zero-flux boundary condition

$$\frac{\partial \left(D_u \nabla u - \chi_c \varphi(u) \nabla c\right)}{\partial \nu} = \frac{\partial c}{\partial \nu} = 0 \qquad \text{on} \quad \partial\Omega \times (0, +\infty), \tag{6.3}$$

where ν is the outward normal vector to the boundary. We assume in the following that the initial condition satisfies

$$\begin{cases} \{u(0,x), c(0,x)\} = \{u^0, c^0\} \in H^1(\Omega) \times H^1(\Omega), \\ 0 \le u^0 \le 1 \text{ a.e. in } \Omega, \quad 0 \le c^0 \le \bar{c} \text{ a.e. in } \Omega, \end{cases} \tag{6.4}$$

where $\bar{c}$ is a positive finite constant.

In the model (6.1)–(6.4), the cell density $u(t,x)$ is attracted by the chemo-attractant given by $c(t,x)$, its concentration. Cells can move randomly by diffusion (where $D_u \ge 0$ is the diffusion coefficient), and by chemotaxis with $\chi_c \ge 0$ a coefficient used to represent the strength of the chemoattraction. A small parameter $\tau > 0$ is used to denote how fast the chemo-attractant is diffusing compared to the cells. Without a loss a generality, we will assume in the following that $\tau = 1$.

We denote by $\varphi(u)$ the chemosensitivity and it is given by

$$\varphi(u) = u(1-u) \qquad \text{for} \qquad 0 \le u \le 1. \tag{6.5}$$

This particular form of chemosensitivity prevents the unrealistic scenario of overcrowding of cells and therefore the blow-up of the solution. Due to this possible behavior of solutions, the Keller-Segel system exhibits a very interesting mathematical structure and the interested reader can refer to the review [117] and the work of Blanchet *et al.* [33]. The volume-filling strategy was proposed in [162] to take into account the finite size of individual cells, leading to the form (6.5).

The Keller-Segel model (6.1)–(6.2) with chemosensitivity (6.5) has a gradient flow structure with the associated energy

$$\mathcal{E}[u,c](t) = \int_\Omega B\left[u\log u - (u-1)\log(1-u)\right] - uc + \frac{1}{2}\left(|\nabla c|^2 + \alpha c^2\right) + C \, \mathrm{d}x, \tag{6.6}$$

where $B = D_u/\chi_c$ and we denote the integral of the nonlinear part of the free energy density by

$$\mathcal{E}_1[u](t) = \int_\Omega F(u) \, \mathrm{d}x,$$

where

$$F(u) = B\left(u\log u - (u-1)\log(1-u)\right) - uc + C. \tag{6.7}$$

Here, C is a positive constant such that $F(u) > 0$, $\forall u \in [0,1]$. For latter convenience, we denote $F'(u) = g(u)$, and we remark that $g'(u) = \frac{1}{\varphi(u)}$. Thus, we can express the Keller-Segel model using its gradient flow structure [34]

$$\partial_t u = \nabla \cdot \left(\chi_c \varphi(u) \nabla \frac{\delta \mathcal{E}}{\delta u}\right), \tag{6.8}$$

$$\tau \partial_t c = -\frac{\delta \mathcal{E}}{\delta c}, \tag{6.9}$$

where the variational derivatives of the energy functional with respect to u and c are given

respectively by

$$\begin{aligned}\frac{\delta \mathcal{E}}{\delta u} &= g(u),\\ \frac{\delta \mathcal{E}}{\delta c} &= -\Delta c + \alpha c - u.\end{aligned}$$

Generally, numerical schemes for a gradient flow model are evaluated by several aspects: *i)* its capacity to keep the energy dissipation; *ii)* if it is convergent, and if error bounds can be established; *iii)* its efficiency; *iv)* its implementation simplicity. For a large class of gradient flows, the Scalar Auxiliary Variable (SAV in short) [180] has shown to fullfil all the previous points. Applying this method to the Keller-Segel model is only possible starting from its gradient flow formulation (6.8)–(6.9) and gives what we call the SAV Keller-Segel model. To define it, we denote by $r(t)$ the scalar auxiliary variable, and set $r(t) = \sqrt{\mathcal{E}_1[u](t)}$ at the continuous level, to get

$$\partial_t u = \nabla \cdot (\chi_c \varphi(u) \nabla \mu_1)\,, \tag{6.10}$$

$$\mu_1 = B\frac{r}{\sqrt{\mathcal{E}_1[u]}} g(u) - c, \tag{6.11}$$

$$\tau \partial_t c = -\mu_2, \tag{6.12}$$

$$\mu_2 = -\Delta c + \alpha c - u. \tag{6.13}$$

We add to this new model an additional equation for r that reads

$$\frac{\mathrm{d}r}{\mathrm{d}t} = \frac{1}{2\sqrt{\mathcal{E}_1[u]}} \int_\Omega g(u) \frac{\partial u}{\partial t}\, \mathrm{d}x. \tag{6.14}$$

In this article, we propose to study a stabilized finite element scheme to simulate the system (6.10)–(6.14) that preserves a modified energy at the discrete level.

Throughout the past decades, the Keller-Segel model as been at the center of many pieces of research. The analytical properties of the Keller-Segel model without volume-filling have been extensively studied. One of the most important result was to show that the solution of the model blows up in finite time if a certain constraint on the initial mass is not satisfied. For the reader interested into the analytical results about this model without volume filling, we refer to the review paper [188]. The volume-filling approach prevents this blow-up of the solution in finite time for any initial condition satisfying (6.4). Moreover, it seems to be more biologically relevant since it takes into account the finite size of the cells. A more general form of the Keller-Segel model is

$$\begin{cases} \partial_t u - \nabla \cdot (D_u \beta(u) \nabla u - \chi_c u \mu(u) \nabla c) = 0,\\ \partial_t c - D_c \Delta c = \delta u - \alpha c, \end{cases} \tag{6.15}$$

where the random movement of the cells is given by $D_u\beta(u)$ (that can be non-linear) and the chemosensitivity is given by $\chi_c \mu(u)$. Particular assumptions on both $\beta(u)$ and $\mu(u)$ can be made to prevent the blow-up of solutions in finite time. The introduction of the parabolic-parabolic KS with volume-filling and quorum-sensing is presented in the work of Painter and Hillen [162]. They described a discrete lattice model where the probability for cells to jump to a different location is dependent on the local density and on the concentration of the chemotactic agent. Their model takes into account the fact that cells may be already present in the neighboring locations. Therefore, the chemotactic movement is set to zero in zones that are too overcrowded. From this discrete model, Painter and Hillen derived the continuous limit model and gave the

following conditions for $\beta(u)$ and $\mu(u)$

$$\beta(u) := \psi(u) - u\psi'(u), \quad \mu(u) \equiv \psi(u),$$

where $\psi(u)$ is a monotonically decreasing function and with the assumptions

$$\psi(0) > 0, \quad \psi(u) > 0 \quad \text{for} \quad 0 < u < u_{\max}, \quad \text{and} \quad \psi(u_{\max}) = 0,$$

where $u_{\max}$ represents the cell density at which chemoattraction stops. The same authors proved the global existence of classical solutions in [116]. They also presented some numerical simulations where they were able to make observation of the behavior of the solution for longer times since the blow-up of the solution is prevented by the model. Many other variations of the Keller-Segel model have been proposed to take into account the effect of volume filling. For example, more recently, Bubba *et al.* [45] proposed to take

$$\psi(u) = \exp\left(-\frac{u}{u_{\max}}\right).$$

Therefore, in the present work we use the chemosensitivity function (6.5) where the maximum cell-density is $\overline{u} = 1$.

Numerical methods for the Keller-Segel model are numerous. Considering zero-flux boundary conditions, the conservation of the total mass of the cells, the positivity of the solution and the capacity to retrieve the energy at the discrete level are the key properties expected from a numerical scheme. Without indication of the contrary, the following works that we review here are about the original form of the Keller-Segel model i.e. $\varphi(u) = u$.

For the parabolic-elliptic Keller-Segel equation where the equation for the chemo-attractant is given by

$$-\Delta c = \delta u - \alpha c,$$

Saito and Suzuki proposed a conservative finite-difference scheme [175]. For the parabolic-parabolic version, Saito proposed for an upwind finite element scheme [173, 172] based on Baba and Tabata's method [19] and performed an error analysis [174]. Other methods have been designed to stabilize the finite element method (i.e. to preserve the positivity of the cell density). The discontinuous Galerkin method has been used and analyzed for the KS model [79, 78, 133], and these schemes have shown very good results in terms of positivity preserving and spatial accuracy. For a standard continuous Galerkin discretization, Strehl *et. al.* [185] used the flux corrected transport (FCT) method [129] to stabilize the scheme and recover an accurate scheme. The main idea behind the FCT method is to add an artificial diffusion term to stabilize the scheme and then correct it to recover a second order scheme in zones where the solution is smooth while the artificial diffusion is kept in regions where the gradient of the solution is large. The finite volume method has also been applied for this problem: we can cite the work of Filbet [86] that deals with the classical Patlak-Keller-Segel model (without volume filling).

The main objective of the previous works that we reviewed so far was to design stable and accurate numerical methods for the Keller-Segel model. However, none of the previous focusses on the dissipation of the energy. The work that is the closer to ours is [12] and deals with the case of the volume-filling chemosensitivity. In the latter, the parabolic-elliptic model is used and the authors were able to prove the preservation of the important properties that are the dissipation of the energy and the positivity of u for two implicit nonlinear finite volume schemes. The difference between the two is that one uses the gradient flow structure of the model while the user uses an exponential rewriting inspired by the Scharfetter-Gummel discretization. To

avoid the solving of the nonlinear system, the authors proposed to compute the chemosensitivity explicitly and use the upwind method to preserve the positivity of the cell density. However, the capacity to prove the dissipation of the energy is lost with this more efficient scheme. To the best of our knowledge, there is no work presenting a linear structure-preserving numerical scheme for the model (6.1)–(6.2).

A recent numerical method to simulate gradient flows that ensures that the energy is preserved at the discrete level is the Scalar Auxiliary Variable method (SAV in short) [181, 180]. This method provides a robust framework to simulate gradient flows in an efficient way. Indeed, the computation of the solution of any gradient flow model requires only the solving of two decoupled linear systems at each time step. This method has shown very interesting results for the simulation of the Cahn-Hilliard equation [179] for which the properties concerning the discrete energy and the conservation of the total mass are of main importance. We must stress that the energy recovered by the SAV method is a modified version of the energy of the real system. This is due to the discretization of the equation for the scalar variable. In a recent work of Bouchriti *et al.* [36], the authors showed that the use of the SAV method for the damped wave equation and the Cahn-Hilliard equation leads to the convergence to modified steady states as well. To the best of our knowledge the SAV method has never been applied to the Keller-Segel model. The principal difference with previous works on the SAV method is that the mobility in the first equation of the Keller-Segel system is not constant through time, leading to the necessity to compute at each time step the associated matrix.

Hence, in this article, we propose to use the SAV method to obtain a new model that we discretize in space using the finite element method, and stabilize it using the multidimensional upwind method proposed in [166]. Altogether, we obtain a new way to simulate the parabolic-parabolic Keller-Segel equation with the certitude to be able to retrieve the positivity of the solution and a modified energy at the discrete level. First, we describe the method and explain the strategy to solve the resulting equations. Then, the well-posedness of the scheme is studied. We show the existence of a unique pair of solution that is non-negative and retrieve the expected L^∞ norm. We also show that the initial mass of the cells is conserved. We prove that a modified energy is retrieved at the discrete level which is an inherent property of the SAV method. Lastly, we present numerical simulations in one dimension, and compare our results with standard numerical techniques for this model.

6.2 Numerical scheme

6.2.1 Finite element framework

Let $L^p(\Omega)$, $W^{m,p}(\Omega)$ with $H^m(\Omega) = W^{m,2}(\Omega)$, where $1 \leq p \leq +\infty$ and $m \in \mathbb{N}$, be respectively the usual Lebesgue and Sobolev spaces. The corresponding norms are respectively $||\cdot||_{m,p,\Omega}$, $||\cdot||_{m,\Omega}$ and semi-norms $|\cdot|_{m,p,\Omega}$, $|\cdot|_{m,\Omega}$. We denote $L^p(0,T;V)$ the Bochner spaces i.e. the spaces with values in Sobolev spaces [3]. The norm in these spaces is defined for all function η Bochner measurable by

$$\|\eta\|_{L^p(0,T;V)} = \left(\int_0^T \|\eta\|_V^p \, \mathrm{d}t \right)^{1/p},$$

and

$$\|\eta\|_{L^\infty(0,T;V)} = \operatorname{ess} \sup_{t\in(0,T)} \|\eta\|_V .$$

The standard L^2 inner product is denoted by $(\cdot,\cdot)_\Omega$ and the duality pairing between $(H^1(\Omega))'$ and $H^1(\Omega)$ by $<\cdot,\cdot>_\Omega$.

Let Ω be a polyhedral domain and $\mathcal{T}^h$, $h>0$, be a quasi-uniform mesh of this domain into $\left|\mathcal{T}^h\right|$ disjoint open mesh elements T. Let $h_T := \mathrm{diam}(T)$ and $h = \max_{T\in\mathcal{T}^h} h_T$. Since the mesh is assumed to be quasi-uniform, we know that it is shape-regular and it exists a positive constant C such that

$$h_T \geq Ch, \quad \forall T \in \mathcal{T}^h.$$

Since the domain is assumed to be polyhedral, the discrete domain Ω_h exactly coïncides with the domain Ω. Hence, the closure of the domain can be written as the union of all the mesh elements $\overline{\Omega} = \overline{\Omega}_h = \bigcup_{T\in T^h} \overline{T}$. We assume that the mesh is acute, *i.e.* for $d=2$ the angles of the triangles can not exceed $\frac{\pi}{2}$ and for $d=3$ the angle between two faces of the same tetrahedron can not exceed $\frac{\pi}{2}$. We define by κ_T the minimal perpendicular length of T and $\kappa_h = \min_{T\in\mathcal{T}^h}\kappa_T$. We introduce the P-1 finite element space associated with the mesh $\mathcal{T}^h$

$$V^h := \{\chi \in C(\overline{\Omega}) : \chi\big|_T \in \mathbb{P}^1(T), \quad \forall T \in \mathcal{T}^h\} \subset H^1(\Omega),$$

where $\mathbb{P}^1(T)$ denotes the space of polynomials of order 1 on T. For latter convenience, we indicate the set of nodes of $\mathcal{T}^h$ by J_h and $\{x_j\}_{j=1,\dots,N_h}$ is the set of their coordinates (with $N_h = |J_h|$ is the total number of nodes). We denote by Λ_i the set of nodes connected to the node x_i by an edge and $G_h = \max_{x_i\in J_h}|\Lambda_i|$.

Therefore, $\{\chi_j\}_{j=1,\dots,N_h}$ is the standard Lagrangian basis functions associated with the spatial mesh. The standard interpolation operator is defined by $\pi^h : C(\overline{\Omega}) \to V^h$ such that

$$\pi^h v(x) = \sum_{i=1}^{N_h} v(x_i)\chi_i(x), \quad \forall v \in C^0(\overline{\Omega}).$$

We denote by $P_h : L^2(\Omega) \to V^h$ the L^2 projection operator

$$(P_h v, \chi) = (v, \chi) \quad \forall v \in L^2(\Omega) \text{ and } \forall \chi \in V^h.$$

We define by $\hat{\chi}_i \in L^\infty(\Omega)$, the characteristic function of the barycentric domain D_i associated with each node x_i (for $i = 1,\dots,N_h$). The barycentric domain D_i is defined as

$$D_i := \bigcup_k \{D_i^k;\, T_k \in \mathcal{T}^h \text{ such that } x_i \in T^k\},$$

and

$$D_i^k := \bigcap_{j=1}^{n_T} \{x; x \in T_k \text{ and } \lambda_{ij}(x) \leq \lambda_i\}$$

where n_T being the number of nodes in the triangle T_k, and λ_i, λ_{ij} (for $j = 1,\dots,n_T$) are the barycentric coordinates with respect to the vertices of the triangles x_i, x_j.

We define the lumped space $\hat{V}_h$ as

$$\hat{V}_h := \{\hat{\chi} : \text{ piecewise constant over barycentric domains i.e. } \hat{\chi}(x) = \hat{\chi}(x_i),\, \forall x \in D_i\}.$$

Therefore, we let the functions $\{\hat{\chi}_j\}_{j=1,\dots,N_h}$ be a basis of the space $\hat{V}_h$ and they are associative with the functions $\{\chi_j\}_{j=1,\dots,N_h}$ i.e. $\chi(x_i) = \hat{\chi}(x_i)$ for all $x_i \in J_h$. We also define the operator

$\hat{\pi}^h : C^0(\overline{\Omega}) \to \hat{V}^h$ given by

$$\hat{\pi}^h v(x) = \sum_{i=1}^{N_h} v(x_i)\hat{\chi}_i(x), \quad \forall v \in C^0(\overline{\Omega}).$$

We define the lumped scalar product by

$$(v_1, v_2)^h = \int_\Omega \pi^h \left(v_1(x)v_2(x)\right) \mathrm{d}x = \left(\hat{v}_1, \hat{v}_2\right), \quad \forall v_1, v_2 \in C^0(\overline{\Omega}),$$

and $\hat{v}_1 = \hat{\pi}^h v_1$. For latter convenience, we state here some well-known results for the P-1 finite element method (see for e.g. [43], [168])

$$|\chi|_{m,p_2} \leq Ch^{-d\left(\frac{1}{p_1}-\frac{1}{p_2}\right)} |\chi|_{m,p_1} \quad \forall \chi \in V^h, 1 \leq p_1 \leq p_2 \leq +\infty, m = 0, 1; \tag{6.16}$$

$$\lim_{h\to 0} \left\|v - \pi^h(v)\right\|_{0,\infty} = 0 \quad \forall v \in C(\overline{\Omega}), \tag{6.17}$$

$$|v - P_h v|_0 + h\,|v - P_h v|_1 \leq Ch^m \|v\|_m \quad v \in H^m(\Omega), \quad m = 1, 2. \tag{6.18}$$

Then from [192] (Lemma 15.1) and [94], we know $\forall v_1, v_2 \in V_h$

$$\left|(v_1, v_2)^h - (v_1, v_2)\right| \leq Ch^2 \|\nabla v_1\|_0 \|\nabla v_2\|_0, \tag{6.19}$$

$$c_1 \|v_1\|_0 \leq \left((v_1, v_1)^h\right)^{1/2} \leq c_2 \|v_1\|_0. \tag{6.20}$$

We define the standard mass M and stiffness K finite element matrices

$$M_{ij} = \int_\Omega \chi_i \chi_j \,\mathrm{d}x, \quad \text{for } i, j = 1, \ldots, N_h,$$

$$K_{ij} = \int_\Omega \nabla\chi_i \nabla\chi_j \,\mathrm{d}x, \quad \text{for } i, j = 1, \ldots, N_h.$$

The lumped mass matrix is a diagonal matrix defined by

$$M_{l,ij} := \int_\Omega \hat{\chi}_i \hat{\chi}_j \,\mathrm{d}x, \quad \text{for } i, j = 1, \ldots, N_h.$$

From the hypothesis we made on the acuteness of the triangulation, we know that (see [94])

$$(\nabla\chi_i, \nabla\chi_j) \leq 0, \quad \text{for } i \neq j. \tag{6.21}$$

Therefore, we know that the non-diagonal entries of the stiffness matrix K and of the matrix A defined below by the equation (6.28) are non-positive.

6.2.2 Fully discrete scheme

Given $N_T \in \mathbb{N}^*$, let $\Delta t := T/N_T$ be the constant time-step and $t^n := n\Delta t$, for $n = 0, \ldots, N_T - 1$. We consider a partitioning of the time interval $[0, T] = \bigcup_{n=0}^{N_T-1} [t^n, t^{n+1}]$. We approximate the continuous time derivative using a forward Euler method $\frac{\partial u_h}{\partial t} \approx \frac{u_h^{n+1} - u_h^n}{\Delta t}$. The finite element numerical problem associated with the system (6.10)–(6.14) is:

Find $\{u_h^{n+1}, c_h^{n+1}\} \in V^h \times V^h$ such that $\forall \chi \in V^h$

$$\left(\frac{u_h^{n+1}-u_h^n}{\Delta t}, \chi\right)^h = -\chi_c \left(\varphi(u_h^n)\nabla \mu_{1,h}^{n+1}, \nabla \chi\right), \tag{6.22}$$

$$\left(\frac{c_h^{n+1}-c_h^n}{\Delta t}, \chi\right) = -\left(\mu_{2,h}^{n+1}, \chi\right), \tag{6.23}$$

$$\left(\mu_{1,h}^{n+1}, \chi\right)^h = -\left(c_h^n, \chi\right)^h + B\left(\frac{g\left(u_h^n\right)}{\sqrt{\mathcal{E}_1[u_h^n]}}, \chi\right)^h r^{n+1}, \tag{6.24}$$

$$\left(\mu_{2,h}^{n+1}, \chi\right) = \left(\nabla c_h^{n+1}, \nabla \chi\right) + \alpha\left(c_h^{n+1}, \chi\right) - \left(u_h^{n+1}, \chi\right)^h, \tag{6.25}$$

$$r^{n+1} - r^n = \frac{1}{2}\left(\frac{g\left(u_h^n\right)}{\sqrt{\mathcal{E}_1[u_h^n]}}, \left(u_h^{n+1}-u_h^n\right)\right)^h, \tag{6.26}$$

where $u_h^n(x) = \sum_{j=1}^{N_h} u_j^n \chi_j(x)$, and $c_h^n(x) = \sum_{j=1}^{N_h} c_j^n \chi_j(x)$ are respectively the finite element approximations of the cell density u, and the concentration of the chemo-attractant c. We add to this system the following initial conditions

$$\begin{cases} \{u_h^0, c_h^0\} = \{\pi^h u^0, \pi^h c^0\} & \text{if } d = 1, \\ \{u_h^0, c_h^0\} = \{P_h u^0, P_h c^0\} & \text{if } d = 2, 3. \end{cases} \tag{6.27}$$

6.2.3 Matrix formulation

Let us define A the finite element matrix associated with the right-hand side of (6.22)

$$A_{ij}^n = \int_\Omega \varphi(u_h^n)\nabla\chi_i \nabla\chi_j \, \mathrm{d}x \quad \text{for } i, j = 1, \ldots, N_h. \tag{6.28}$$

and the variable

$$s_{1,h}^n = \frac{g_h^n}{\sqrt{\mathcal{E}_1[u_h^n]}},$$

where $g_h^n(x_i) = g(u_h^n(x_i))$ for all $x_i \in J$. We denote in the following by capital letters the vectors associated with the quantities denoted by small letters in the finite element problem. Therefore, the system (6.22)–(6.26) can be rewritten into a matrix formulation

$$\begin{cases} M_l \frac{U^{n+1}-U^n}{\Delta t} & = -\chi_c A^n W_1^{n+1}, \\ M \frac{C^{n+1}-C^n}{\Delta t} & = -M W_2^{n+1}, \\ W_1^{n+1} & = -C^n + B S_1^n r^{n+1}, \\ M W_2^{n+1} & = K C^{n+1} + \alpha M C^{n+1} - M_l U^{n+1}, \\ r^{n+1} - r^n & = \frac{1}{2}\left(S_1^n\right)^T M_l \left(U^{n+1} - U^n\right). \end{cases}$$

However the simulation of the problem (6.22)–(6.26) by the standard finite element method is well-known to produce to non-physical solutions. Due to the advection term associated to the effect of chemotaxis, when the ratio χ_c/D_n becomes too large, the solution is expected to oscillate with the apparition of upper and undershoots for u_h^n. To avoid this issue, we propose to compute the matrix associated with advection using the upwind method.

6.2.4 Upwind stabilization

The key idea behind the upwind method of finite elements proposed in [166] is to modify each entry of the matrix A^n defined by (6.28) such that

$$\left(\overline{A}^n\right)_{ij} = \overline{\varphi}^n_{\mathrm{ij}} K_{ij}. \tag{6.29}$$

The coefficient $\overline{\varphi}^n_{\mathrm{ij}}$ is a constant (at time t^n) along the edge connecting the nodes i and j. Therefore, for each edge, we evaluate the sign of the gradient ∇c^n_h and compute

$$(\overline{\varphi}^n)_{ij} = \begin{cases} U^n_j\left(1-U^n_i\right) & \text{if} \quad C^n_i - C^n_j - B(g^n_i - g^n_j) > 0, \\ U^n_i\left(1-U^n_j\right) & \text{otherwise.} \end{cases} \tag{6.30}$$

Therefore, our new problem now reads: Find $\{u^{n+1}_h, c^{n+1}_h, r^{n+1}\} \in V^h \times V^h \times \mathbb{R}$ such that $\forall \chi \in V^h$

$$\left(\frac{u^{n+1}_h - u^n_h}{\Delta t}, \chi\right)^h = -\chi_c \left(\overline{\varphi}(u^n_h)\nabla \mu^{n+1}_{1,h}, \nabla\chi\right), \tag{6.31}$$

$$\left(\frac{c^{n+1}_h - c^n_h}{\Delta t}, \chi\right) = -\left(\mu^{n+1}_{2,h}, \chi\right), \tag{6.32}$$

$$\left(\mu^{n+1}_{1,h}, \chi\right)^h = -\left(c^n_h, \chi\right)^h + B\left(s^n_{1,h}, \chi\right)^h r^{n+1}, \tag{6.33}$$

$$\left(\mu^{n+1}_{2,h}, \chi\right) = \left(\nabla c^{n+1}_h, \nabla\chi\right) + \alpha\left(c^{n+1}_h, \chi\right) - \left(u^{n+1}_h, \chi\right)^h, \tag{6.34}$$

$$r^{n+1} - r^n = \frac{1}{2}\left(s^n_{1,h}, (u^{n+1}_h - u^n_h)\right)^h, \tag{6.35}$$

and its matrix formulation is

$$M_l \frac{U^{n+1} - U^n}{\Delta t} = -\chi_c \overline{A}^n W^{n+1}_1, \tag{6.36}$$

$$M \frac{C^{n+1} - C^n}{\Delta t} = -M W^{n+1}_2, \tag{6.37}$$

$$W^{n+1}_1 = -C^n + B S^n_1 r^{n+1}, \tag{6.38}$$

$$M W^{n+1}_2 = K C^{n+1} + \alpha M C^{n+1} - M_l U^{n+1}, \tag{6.39}$$

$$r^{n+1} - r^n = \frac{1}{2}\left(S^n_1\right)^T M_l \left(U^{n+1} - U^n\right). \tag{6.40}$$

6.2.5 Solving Algorithm

To solve the solution of the system (6.36)–(6.40) while avoiding to invert the matrix $\overline{A}^n$ at each time step, we use the decomposition $U^{n+1} = U^{n+1}_1 + r^{n+1}U^{n+1}_2$ and solve successively the following set of equations

$$U^{n+1}_1 = \Delta t \chi_c M_l^{-1}\left(\overline{A}^n C^n\right) + U^n, \tag{6.41}$$

$$U^{n+1}_2 = -\Delta t D_u M_l^{-1}\left(\overline{A}^n S^n_1\right), \tag{6.42}$$

$$r^{n+1} = \frac{r^n + \frac{1}{2}\left(S_1^n\right)^T M_l \left(U_1^{n+1} - U^n\right)}{1 - \frac{1}{2}\left(S_1^n\right)^T M_l U_2^{n+1}}, \tag{6.43}$$

$$U^{n+1} = U_1^{n+1} + r^{n+1} U_2^{n+1}, \tag{6.44}$$

$$C^{n+1} = \left((1+\Delta t\alpha)M + \Delta t K\right)^{-1}\left(M\left(\Delta t U^{n+1} + C^n\right)\right). \tag{6.45}$$

This algorithm is well suited for problems involving non-constant mobility matrices since only constant matrices are inverted only once at the beginning of the simulation.

6.3 Existence of a non-negative solution and stability bound

6.3.1 Existence of a discrete non-negative solution

Theorem 37 (Existence of a unique non-negative discrete solution) *Let $d \leq 3$ and assume that $\kappa_h > 0$, $\Delta t > 0$, $D_u > 0$ such that*

$$\frac{(d+1)\, G_h\, \Delta t\, \chi_c}{D_n\, r^{n+1}\, \kappa_h^2} \max_{\substack{i=1,\dots,N_h \\ j\in\Lambda_i}} \left|C_i^n - C_j^n\right| \leq 1, \tag{6.46}$$

and, given an initial condition $\{u_h^0, c_h^0\}$ such that (6.4) and (6.27) are satisfied, the problem (6.31)–(6.35) admits a unique solution $\{u_h^{n+1}, c_h^{n+1}, r^{n+1}\} \in V^h \times V^h \times \mathbb{R}$ with

$$0 \leq u_h^{n+1} \leq 1, \quad \text{and} \quad 0 \leq c_h^{n+1} \leq c_{max},$$

where c_{max} is a positive and finite constant.

Proof.

Step 1: Existence of a unique solution in $V^h \times V^h \times \mathbb{R}$. As we have seen in the section describing the numerical scheme, the problem (6.31)–(6.35) reduces to the solving of five decoupled equations (6.41)–(6.45). From equation (6.43), and the fact that $\mathcal{E}_1[u^n]$ is defined up to a constant C_0 that can be defined arbitrary. One must be careful choosing this constant such that r^{n+1} is well defined. Then, the existence and uniqueness of a solution $\{u_h^{n+1}, c_h^{n+1}, r^{n+1}\} \in V_h \times V_h \times \mathbb{R}$ follows from the Lax-Milgram theorem and since $\{u_h^n, c_h^n, r^n\}$ are known and bounded.

Step 2: Conservation of mass. To prove mass conservation, we use the identity

$$\sum_{\substack{j\neq i \\ x_j\in T_i}} \left|\overline{A}_{ij}^n\right| = \overline{A}_{ii}^n. \tag{6.47}$$

Therefore, for each $x_i \in J_h$, we have

$$\sum_{j=1}^{N_h} \left(\hat{\chi}_j, \hat{\chi}_i\right)\left(u_h^{n+1} - u_h^n\right)(x_j) = \Delta t \left[-D_u r^{n+1} \sum_{j=1}^{N_h} \overline{A}_{ij}^n s_{1,h}^n(x_j) + \chi_c \sum_{j=1}^{N_h} \overline{A}_{ij}^n c_h^n(x_j)\right].$$

Summing over the nodes, we get

$$\sum_{i=1}^{N_h}\sum_{j=1}^{N_h}(\hat{\chi}_j,\hat{\chi}_i)\left(u_h^{n+1}-u_h^n\right)(x_j)$$
$$=\Delta t\left[-D_u r^{n+1}\sum_{i=1}^{N_h}\sum_{j=1}^{N_h}\overline{A}_{ij}^n s_{1,h}^n(x_j)+\chi_c\sum_{i=1}^{N_h}\sum_{j=1}^{N_h}\overline{A}_{ij}^n c_h^n(x_j)\right].$$

Using the symmetry of the matrix A, the property (6.47) and the fact that the mesh is acute, we obtain

$$\sum_{i=1}^{N_h}\sum_{j=1}^{N_h}(\hat{\chi}_j,\hat{\chi}_i)\left(u_h^{n+1}-u_h^n\right)(x_j)=0,$$

which implies mass conservation .

Step 3: Positivity and L^∞ bound for $\{u_h^{n+1},c_h^{n+1}\}$. It is well known that the loss of the positivity (and the eventual uppershoot) of the solution is associated with the advection term. The diffusion term on the other hand regularizes the solution and does not lead to instabilities. Therefore, to examinate the stability of our scheme, we set $D_u=0$. Therefore, in this case, U^{n+1} is given by the equation (6.41) (since r^{n+1} does not appear anymore in (6.33)). Then, we have for each node $x_i\in J_h$

$$U_i^{n+1}=U_i^n+\frac{\Delta t\chi_c}{(M_l)_{ii}}\sum_{x_j\in\Lambda_i}\overline{\varphi}_{ij}^n K_{ij}(C_j^n-C_i^n).$$

From (6.21) and the definition of the upwind chemosensitivity (6.30), we know that we could lose the positivity of the cell density if $C_j^n>C_i^n$. Thus, to preserve it, we must ensure the condition

$$U_i^n+\frac{\Delta t\chi_c}{(M_l)_{ii}}\sum_{x_j\in\Lambda_i}U_i^n\left(1-U_j^n\right)K_{ij}\max\left(0,C_j^n-C_i^n\right)\geq 0.$$

The same holds to preserve $\left\|u_h^{n+1}\right\|_{L^\infty}\leq 1$, we recover the condition

$$U_i^n+\frac{\Delta t\chi_c}{(M_l)_{ii}}\sum_{x_j\in\Lambda_i}U_j^n\left(1-U_i^n\right)K_{ij}\min\left(0,C_j^n-C_i^n\right)\leq 1.$$

However, we know that [94],

$$\frac{|K_{ij}|}{M_{l,ii}}\leq\frac{K_{ii}}{M_{l,ii}}\leq\frac{(d+1)}{\kappa_h^2}.$$

Therefore, to preserve the physical bound of the solution we must ensure that the condition (6.46) is satisfied. Then, knowing that $0\leq u_h^{n+1}\leq 1$, the non-negativity and the existence of an upper bound $c_{\max}$ such that

$$0\leq c_h^{n+1}\leq c_{\max},$$

is trivially found from the properties of M-matrices. This finishes the proof of the existence of the solution of the problem (6.22)–(6.27). □

6.3.2 Discrete energy a priori estimate

Since we are using the SAV method, we can compute a modified version of the energy at the discrete level.

Proposition 38 (Discrete energy) *Consider a solution* $\{u_h^{n+1}, c_h^{n+1}\}$ *defined by Theorem 37, the discrete energy of the system* (6.22)–(6.26) *is given by*

$$E(u_h^{n+1}, c_h^{n+1}) = \frac{1}{2}\left(\left|c_h^{n+1}\right|_1^2 + \alpha\left\|c_h^{n+1}\right\|_0^2\right) + B\left|r^{n+1}\right|^2 - \left(c_h^{n+1}, u_h^{n+1}\right)^h, \tag{6.48}$$

and

$$\frac{\mathrm{d}E}{\mathrm{d}t} := \frac{E^{n+1} - E^n}{\Delta t} \leq -\left(\left\|\mu_{2,h}^{n+1}\right\|_0^2 + \chi_c \int_\Omega \overline{\varphi}(u_h^n)\left|\nabla \mu_{1,h}^{n+1}\right|^2 \mathrm{d}x\right). \tag{6.49}$$

Proof. Starting from equation (6.22) with $\chi = \mu_{1,h}^{n+1}$, we have

$$\left(u_h^{n+1} - u_h^n, \mu_{1,h}^{n+1}\right)^h = -\Delta t \chi_c \int_\Omega \overline{\varphi}(u_h^n)\left|\nabla \mu_{1,h}^{n+1}\right|^2 \mathrm{d}x.$$

The same can be done starting from equation (6.23) to obtain

$$\left(c_h^{n+1} - c_h^n, \mu_{2,h}^{n+1}\right) = -\Delta t \left\|\mu_{2,h}^{n+1}\right\|_0^2.$$

Therefore, summing the two previous equations, we obtain

$$\left(u_h^{n+1} - u_n^h, \mu_{1,h}^{n+1}\right)^h + \left(c_h^{n+1} - c_n^h, \mu_{2,h}^{n+1}\right) = -\Delta t\left(\left\|\mu_{2,h}^{n+1}\right\|_0^2 + \chi_c \int_\Omega \overline{\varphi}(u_h^n)\left|\nabla \mu_{1,h}^{n+1}\right|^2 \mathrm{d}x\right),$$

from which we conclude (6.49). Consequently, we already recover the monotonic decay of the discrete energy. To obtain the expression of the energy, we replace $\chi = u_h^{n+1} - u_h^n$ in (6.24) to get

$$\left(u_h^{n+1} - u_h^n, \mu_{1,h}^{n+1}\right)^h = -\left(u_h^{n+1} - u_h^n, c_h^n\right)^h + Br^{n+1}\left(u_h^{n+1} - u_h^n, s_{1,h}^n\right)^h.$$

However, using the equation (6.26), we have

$$\left(u_h^{n+1} - u_h^n, \mu_{1,h}^{n+1}\right)^h = -\left(u_h^{n+1} - u_h^n, c_h^n\right)^h + 2Br^{n+1}\left(r^{n+1} - r^n\right).$$

Moreover, using the inequality $a(a-b) \geq \frac{1}{2}\left(a^2 - b^2\right)$, we get

$$\left(u_h^{n+1} - u_n^h, \mu_{1,h}^{n+1}\right)^h \geq -\left(c_h^n, u_h^{n+1} - u_n^h\right)^h + B\left|r_1^{n+1}\right|^2 - B\left|r_1^n\right|^2. \tag{6.50}$$

Then, performing the same calculations starting from the equation (6.25), we obtain

$$\left(c_h^{n+1} - c_n^h, \mu_{2,h}^{n+1}\right) \geq \frac{1}{2}\left[\left|c_h^{n+1}\right|_1^2 - \left|c_h^n\right|_1^2 + \alpha\left(\left\|c_h^{n+1}\right\|_0^2 - \left\|c_h^n\right\|_0^2\right)\right] - \left(u_h^{n+1}, c_h^{n+1} - c_h^n\right)^h. \tag{6.51}$$

Summing equation (6.50) with (6.51), we obtain the inequality

$$\begin{aligned}\frac{1}{2}\left[\left|c_h^{n+1}\right|_1^2 - \left|c_h^n\right|_1^2 + \alpha\left(\left\|c_h^{n+1}\right\|_0^2 - \left\|c_h^n\right\|_0^2\right)\right] + B\left|r_1^{n+1}\right|^2 - B\left|r_1^n\right|^2 - \left(u_h^{n+1}, c_h^{n+1}\right)^h + \left(u_h^n, c_h^n\right)^h \\ \leq -\Delta t\left(\left\|\mu_{2,h}^{n+1}\right\|_0^2 + \chi_c \int_\Omega \overline{\varphi}(u_h^n)\left|\nabla \mu_{1,h}^{n+1}\right|^2 \mathrm{d}x\right),\end{aligned}$$

from which we deduce the definition and the decay of the discrete energy (6.48)–(6.49). □

Parameters	Stable	Unstable
Δt	0.001	0.0001
Δx	0.1	0.1
u^0	0.5	0.5
χ_c	40	120
D_u	1	1
α	0.5	0.5
τ	0.01	0.01
C_{SAV}	10^{10}	10^{10}

Table 6.1 – Parameters of the 1D test cases

Remark 39 *From the fact that both u_h^{n+1} and c_h^{n+1} are bounded (see Theorem 37), the energy defined by (6.48) is bounded from below.*

6.4 Numerical results

We now present the numerical results obtained by the upwind finite element SAV scheme for the Keller-Segel model. We first show the dissipation of the discrete energy and the conservation properties of our scheme in a simple 1D test case. We use this test case to compare the SAV scheme with other methods. Interestingly, we observe that the SAV-upwind scheme enhances the spatial accuracy compared to a simple upwind scheme. To verify this effect, we analyze the convergence of the scheme numerically for different h.

6.4.1 1D numerical results

Stable and unstable test cases

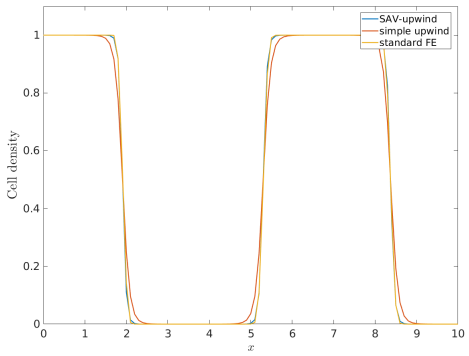
We consider two test cases: in the first we consider a choice of parameter such that even the consistent finite element method is stable. This standard method gives us our reference high-order solution. Computing for the same parameters and initial conditions the solution given by a classical upwind method, and our SAV-upwind scheme, we are able to discuss the spatial accuracy of our method in the stable regime. Then, we consider a second test case such that the standard finite element scheme is unstable. The two upwind methods are expected to remain stable for this choice of parameter, but we will discuss the amount of diffusion added by both schemes.

Table 6.1 summarizes the parameters used in the one dimensional numerical schemes for the stable and the unstable test cases. The parameter C_{SAV} in Table 6.1 corresponds to the constant that needs to be taken in the energy so that the energy remains positive. We choose a very large value to be sure that we will not have issues coming from the calculation of $\sqrt{\mathcal{E}_1(u_h^n)}$. We precise that this value can also be modified at each time step. Indeed, knowing the value of $\mathcal{E}_1(u_h^n)$, we choose C_{SAV} such that $\mathcal{E}_1(u_h^n) + C_{\text{SAV}} > 0$. The initial cell density is a uniform distributed random perturbation around the value u^0.

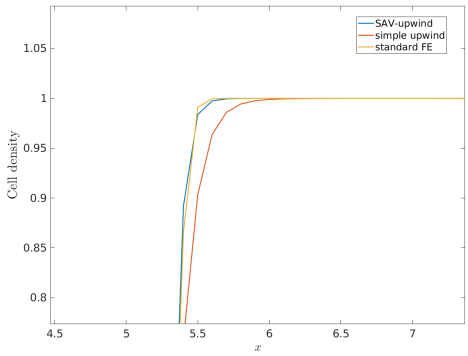
Figure 6.1a shows the solution u_h given by the three different schemes at time $T = 5$ for the stable case. We observe that the three schemes reach a meta-stable state with well-delineated aggregates formed in the domain. Figure 6.1b is a zoom of the previous figure on the top of an aggregate. That way, we clearly observe the discrepancies between our reference solution given by the standard P-1 finite element method, and the two upwind schemes. The blue curve

corresponding to the solution of the SAV-upwind method is really close to the reference solution (yellow curve) while the classical upwind method (red curve) fails to represent the sharpness of the aggregates. Indeed, in this region, the solution is too smooth for the upwind scheme.

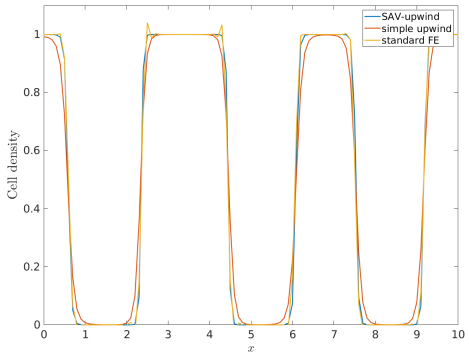
Figure 6.1c depicts the aggregates formed at time $T = 5$ by the three schemes for the unstable case. Again, the aggregates are located at the same positions for the three schemes. We see that the solution given by the standard finite element discretization oscillates in zones of large variations. Zooming at the top of an aggregate (Figure 6.1d), we clearly see that effect. However, the two upwind methods remain stable and do not oscillate. We observe that the SAV-upwind method gives a solution that seems to correspond to the average of the yellow curve. Indeed, the SAV-upwind method gives a solution that is sharper than the classical upwind method. This result indicates that the SAV method allows to retrieve a higher order solution compared with a classical upwind scheme while remaining stable.



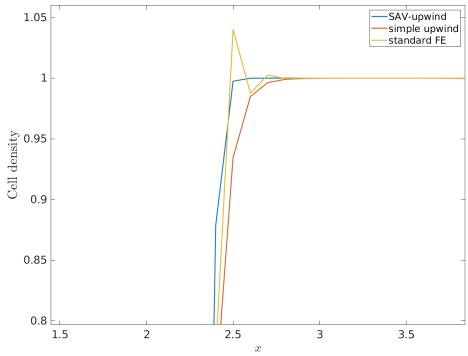
(a) Stable case



(b) Top of an aggregate for stable case



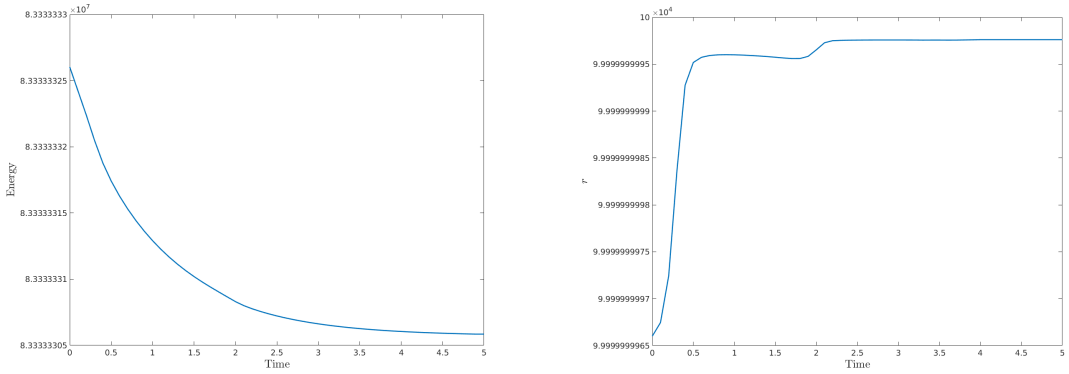
(c) Unstable case



(d) Top of an aggregate for unstable case

Figure 6.1 – End state of the simulations of the volume-filling KS model for a standard finite element discretization, a classical upwind method in finite element, the SAV-upwind method.

For the unstable test case, Figure 6.2 shows the evolution of the modified energy and of the scalar variable r given by the SAV-upwind scheme during the simulation. As expected by our calculations, the modified energy decreases monotonically and converges to a plateau. The scalar auxiliary variable r increases during time and reaches a plateau as well.



(a) (b)

Figure 6.2 – Evolution of the modified energy E (left) and the scalar variable r (right) through time.

Parameters	
Δt	10^{-5}
N_h	$[1000, 750, 500, 200]$
u^0	0.5
χ_c	40
D_u	1
α	0.5
τ	0.01
C_{SAV}	10^{10}

Table 6.2 – Parameters of the calculation of the convergence order

Based on the observation, made on these two test cases, we are interested to compute the numerical order of convergence.

Numerical order of convergence

In this section, we use the parameters summarized in Table 6.2.

We compute the error of the SAV-upwind solution using a reference solution at time $T = 5$ computed using the standard finite element scheme on the fine mesh $h = 10^{-2}$ corresponding to 1000 nodes in the domain. Then, we vary the number of nodes $N_h = [1000, 750, 500, 200]$, and for each simulation we compute the L^2 error $\|u_{\text{ref}} - u_h\|_{L^2(\Omega)}$. The results are shown on Figure 6.3. As a reference, we also show on this figure, the two straight lines representing first-order (yellow) and second-order (red) convergence. We see that the spatial order of convergence for our SAV-upwind method is between first and second-order. Knowing that a classical upwind method is at most-first order accurate, this result indicates that the SAV method enhances the spatially accuracy.

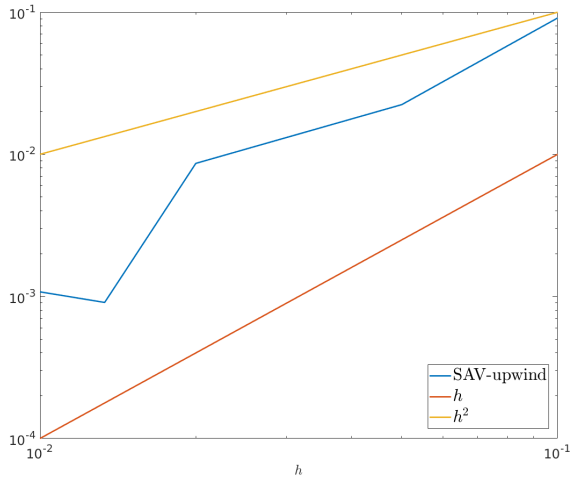


Figure 6.3 – L^2 norm of the error in function of the number of nodes at time $T = 5$.

6.5 Conclusion

We presented the application of Scalar Auxiliary Variable method to the parabolic-parabolic Keller-Segel with volume filling using the gradient flow structure of the model. The resulting equations were approximated using a simple $P-1$ finite element method, and stabilized by a multi-dimensional upwind method. The system is composed of linear coupled equations that can be solved efficiently using a decomposition of the solution. We were able to prove for this system the existence of a unique positive and bounded solution that the preserves the monotonic decay of the discrete energy. We must stress that from the use of the SAV method, the energy that we are able to recover is a modified version of the standard one. Finally, we presented numerical simulations that indicate a better accuracy in space of our method compared to a classical upwind method.

Chapter 7

Conservation properties and long time behavior of the Scalar Auxiliary Variable method for nonlinear dispersive equations.

Abstract

We carry out the convergence analysis of the Scalar Auxiliary Variable (SAV) method applied to the nonlinear Schrödinger equation which preserves a modified Hamiltonian on the discrete level. We derive a weak and strong convergence result, establish second-order global error bounds and present long time error estimates on the modified Hamiltonian. In addition, we illustrate the favorable energy conservation of the SAV method compared to classical splitting schemes in certain applications.

This chapter is taken from A. P., Katharina Schratz *Convergence, error analysis and longtime behavior of the Scalar Auxiliary Variable method for the nonlinear Schrödinger equation*, Submitted, (2020).

7.1 Introduction

We consider the Gross-Pitaevskii [107] equation (NLS) set on the d-dimensional torus $\Omega = \mathbb{T}^d = (\mathbb{R}/2\pi\mathbb{Z})^d$ (where $d \leq 3$)

$$i\partial_t u(t,x) = -\Delta u(t,x) + V(x)u(t,x) + f\left(|u(t,x)|^2\right) u(t,x), \quad t \in (0,T] \tag{7.1}$$

with initial conditions $u(0,x) = u^0(x)$, a real-valued interaction potential $V(x)$ and nonlinearity $f(|u|^2)$.

The Hamiltonian energy associated to equation (7.1) takes the form

$$H(u,\overline{u}) = \frac{1}{2}\int_\Omega \left(|\nabla u|^2 + V(x)\,|u|^2 + F\left(|u|^2\right)\right) \mathrm{d}x,$$

where $F\left(|u|^2\right)$ is defined by $F'\left(|u|^2\right) = f\left(|u|^2\right)$. Note that the Hamiltonian $H(u(t),\overline{u}(t))$ as well as the probability density $\|u(t,\cdot)\|^2_{L^2(\Omega)}$ is preserved by the system (7.1).

In the following we will denote by $\mathcal{E}_1$ the sum of the nonlinear and potential part of the Hamiltonian

$$\mathcal{E}_1 = \frac{1}{2}\int_\Omega V(x)\left|u\right|^2 + F\left(\left|u\right|^2\right)\mathrm{d}x.$$

Using the decomposition $u(t,x) = p(t,x) + iq(t,x)$, equation (7.1) can be furthermore rewritten as the Hamiltonian system

$$\begin{cases} \partial_t p &= -\Delta q + \frac{\delta \mathcal{E}_1[t]}{\delta q}, \\ \partial_t q &= \Delta p - \frac{\delta \mathcal{E}_1[t]}{\delta p}, \end{cases} \tag{7.2}$$

with the associated Hamiltonian

$$H(p,q) = \frac{1}{2}\int_\Omega |\nabla p|^2 + |\nabla q|^2 + V(x)\left(|p|^2 + |q|^2\right) + F\left(|p|^2, |q|^2\right)\mathrm{d}x. \tag{7.3}$$

In this notation, $\mathcal{E}_1$ takes the form

$$\mathcal{E}_1 = \frac{1}{2}\int_\Omega V(x)\left(|p|^2 + |q|^2\right) + F\left(|p|^2, |q|^2\right)\mathrm{d}x.$$

Due to their importance in numerous applications, reaching from Bose-Einstein condensation over nonlinear optics up to plasma physics, nonlinear Schrödinger equations are nowadays very well studied numerically. In the last decades a large variety of different numerical schemes has been proposed [22, 15, 24, 104, 105]. Thanks to their simplicity and accuracy, a popular choice thereby lies in so-called splitting methods, where the right hand side of (7.1) is split into the linear and nonlinear part, respectively, see, e.g., [25, 31, 21] and the references therein. The popularity of splitting methods also stems from their structure preservation. They conserve exactly the L^2 norm of the solution and allow for near energy conservation over long times, see, e.g., [83]. However, in [159] the authors show that in certain applications splitting methods suffer from severe order reduction such as in case of non-linearities with non-integer exponents. The latter arises for instance in context of optical dark and power law solitons with surface plasmonic interactions [65]. As a solution to that issue, the authors proposed in [159] a new class of low regularity exponential-type integrators for NLS. In this article we use a different approach based on the so-called Scalar Auxiliary Variable (SAV) method which was originally proposed to design structure-preserving numerical schemes for gradient flows [181, 180]. Very recently it also became popular in context of Hamiltonian systems [18, 93, 49, 84]. The main advantage of the SAV method lies in the fact that it preserves a modified Hamiltonian on the discrete level. Due to its generality, it can be applied to a large class of equations involving any kind of nonlinearity. The resulting numerical schemes are linearly implicit and allow for efficient calculations.

The main idea behind the SAV method is to introduce a scalar variable $r(t) = \sqrt{\mathcal{E}_1 + \mathcal{E}_c}$ that will become an unknown at the discrete level and where the arbitrary constant $\mathcal{E}_c > 0$ is used to obtain $\mathcal{E}_1 + \mathcal{E}_c > 0$. We must stress that one as to be very careful with the choice of the constant $\mathcal{E}_c$. Indeed, it is well known that even for the cubic non-linearity, *i.e.* $f(|p|^2, |q|^2) = \beta\left|p^2 + q^2\right|^2$ with $\beta < 0$ (focussing NLSE), the hamiltonian energy (7.3) is not bounded from below a priori. In the following analysis, we implicit assume that it exists a constant $\mathcal{E}_c$ such that $\mathcal{E}_1 + \mathcal{E}_c > 0$, which is often the case in the study of Bose-Einstein condensate as pointed out by Antoine *et al.* [18]. In practice, we compute the term $\mathcal{E}_1$ explicitly and therefore one can adapt the constant $\mathcal{E}_c$ during the simulation. The system is supplemented by an equation describing the time evolution of $r(t)$. In case of the nonlinear Schrödinger equation (7.1) the continuous SAV model

takes the form

$$\begin{cases} \partial_t p &= -\Delta q + r(t) g_1(p,q), \\ \partial_t q &= \Delta p - r(t) g_2(p,q), \\ \partial_t r(t) &= \frac{1}{2}\left[(g_1(p,q), \partial_t q) + (g_2(p,q), \partial_t p) \right], \end{cases} \tag{7.4}$$

where $(\cdot,\cdot)$ denotes the standard L^2 scalar product and

$$g_1(p,q) = \frac{1}{\sqrt{\mathcal{E}_1[t] + \mathcal{E}_c}} \frac{\delta \mathcal{E}_1[t]}{\delta q}, \quad g_2(p,q) = \frac{1}{\sqrt{\mathcal{E}_1[t] + \mathcal{E}_c}} \frac{\delta \mathcal{E}_1[t]}{\delta p}.$$

Associated to this SAV model we find the Hamiltonian

$$\tilde{H}(p,q) = \frac{1}{2} \int_\Omega |\nabla p|^2 + |\nabla q|^2 \, \mathrm{d}x + |r|^2,$$

which is conserved by the SAV model (7.4). In the following, we assume that for $i = 1, 2$

$$|g_i'(p,q)| \leq C\left((|p| + |q|)^\beta + 1 \right), \quad |g_i''(p,q)| \leq C\left((|p| + |q|)^{\beta'} + 1 \right), \tag{7.5}$$

for some $\beta, \beta' > 0$.

Remark 40 *In this paper, we focus on the Gross-pitaevskii equation under the form (7.1). Even though the choice of the nonlinearity and, therefore, the precise form of $\mathcal{E}_1$, depends on the structure of the considered Schrödinger equation, we highlight that the SAV scheme is, in its design, general enough to work for a large number of applications. Indeed, as long as there exists a constant $\mathcal{E}_c$ such that for all times $t \geq 0$ it holds that $\mathcal{E}_1 + \mathcal{E}_c > 0$ we can apply the SAV method. Therefore, modifications such as the effect of dipole-dipole interactions, rotating GPE (see Antoine et al. [18]), or even time dependent potentials $V = V(t,x)$ can be taken into account. For an extensive overview on applications and generalisations of the nonlinear Schrödinger equation, we refer the interested reader to the review article of Bao and Cai [22] and the references therein.*

Following the works of Antoine *et al.* [18] and Fu *et al.* [93], we analyze a fully discrete SAV scheme for the nonlinear Schrödinger equation (7.1) based on a Crank-Nicholson time discretization of the NLS SAV model (7.4) coupled with a pseudo-spectral discretization for the spatial discretization. Energy conservation properties of the SAV method for nonlinear Schrödinger equations were recently derived in [18, 93] and their convergence was extensively tested numerically. Very recently, Feng *et al.* [84] use the SAV method to design arbitrary high order space-time finite element scheme for the nonlinear Schrödinger equation. While their method uses a finite element discretization in space, we propose in this work to use a Fourier pseudospectral discretization. The main contribution of this article lies in establishing global error estimates on the fully discrete Fourier-PseudoSpectral Crank-Nicholson NLS SAV scheme (CN-SAV-SP in short). More precisely, we derive weak and strong convergence and prove second order error estimates for the fully discrete scheme. Our theoretical convergence analysis is inspired by the analysis of the SAV method in the context of gradient flows [179]. We underline our convergence results with numerical experiments and compare the SAV scheme with classical splitting methods. Our numerical findings suggest that in certain cases, such as in case of non-linearities involving a non-integer exponent, the SAV scheme preserves its second order energy conservation property while classical splitting methods suffer from sever order reduction. We also conduct numerical experiments showing that the SAV scheme is able to compute correctly ground states of Bose-Einstein condensates.

Outline of the paper. In the first part of the paper, we carry out a fully discrete error analysis of the SAV scheme and establish second order convergence estimates, see Theorem 56. Our theoretical convergence results are then numerically underlined in the second part of the paper, see Section 7.6.

Notations. Let $L^p(\Omega)$, $W^{m,p}(\Omega)$ with $H^m(\Omega) = W^{m,2}(\Omega)$, where $1 \leq p \leq +\infty$ and $m \in \mathbb{N}$, denote the standard Lebesgue and Sobolev spaces equipped with the corresponding norms $||\cdot||_{m,p}$, $||\cdot||_m$ and semi-norms $|\cdot|_{m,p}$, $|\cdot|_m$. We also denote by $H^m_p(\Omega)$ the subset of $H^m(\Omega)$ that consists of 2π-periodic functions that are in $H^m(\Omega)$. We denote by $L^p(0,T;V)$ the Bochner spaces i.e. the spaces with values in Sobolev spaces [3]. The norm in these spaces is defined for all Bochner measurable functions η by

$$\|\eta\|_{L^p(0,T;V)} = \left(\int_0^T \|\eta\|_V^p \, \mathrm{d}t\right)^{1/p}, \qquad \|\eta\|_{L^\infty(0,T;V)} = \operatorname{ess}\sup_{t\in(0,T)} \|\eta\|_V .$$

The standard L^2 inner product is denoted by $(\cdot,\cdot)_\Omega$ and the duality pairing between $\left(H^1(\Omega)\right)' = H^{-1}(\Omega)$ and $H^1(\Omega)$ by $<\cdot,\cdot>_\Omega$. The dual space $H^{-1}(\Omega)$ is endowed with the norm

$$\|\phi\|_{H^{-1}(\Omega)} = \sup_{\eta\in H^1(\Omega)} \{<\phi,\eta>_\Omega, \quad \|\eta\|_1^2 \leq 1\}.$$

Remark 41 *Even though our model problem (7.1) is equipped with periodic boundary conditions, our analysis holds for homogeneous Dirichlet or Neumann boundary conditions.*

7.2 Numerical scheme

7.2.1 Time and space discretisation of the SAV model

We use a standard Fourier pseudospectral method [50, 87, 106] for the spatial discretization of the SAV model (7.4). We refer the reader to the book of Trefethen [195] for details of the implementation of such scheme in MATLAB. We emphasize that our paper presents numerical simulation in dimension $d = 1$. However, the method can be adapted to higher dimensions. Our convergence and error analysis holds in dimensions $1 \leq d \leq 3$.

Thereby, for the sake of clarity, we here give the details of the space discretization for $d = 1$. We denote by X_N the space spanned by the trigonometric functions up to degree $N/2$

$$X_N := \operatorname{span}\{e^{ikx/L} : -N/2 \leq k \leq N/2 - 1\}.$$

For the time discretisation of the SAV system (7.4) we apply a Crank-Nicholson discretisation with time step τ such that $t^k = k\tau$ for $k \in \mathbb{N}$. At each grid point we thereby approximate the time derivative by

$$\partial_t u(t^{k+1}, x) \approx \frac{u(t^{k+1}, x) - u(t^k, x)}{\tau}.$$

Let us give the details of the approximation in dimension $d = 1$, where the domain is defined by $\Omega = [-\pi, \pi]$ with a mesh size h. In this case the collocation points are $x_a = \frac{2\pi a}{N}$ where $a \in \mathcal{B}$ with

$$\mathcal{B} := \begin{cases} \{-P, \ldots, P-1\} & \text{if} \quad N = 2P \quad \text{is even,} \\ \{-P, \ldots, P\} & \text{if} \quad N = 2P+1 \quad \text{is odd.} \end{cases}$$

We denote by $U^k(x_a)$ the approximation of $u(t^k, x_a)$. The Fourier pseudo-spectral discretization is given by

$$U^k(x_a) = \sum_{p\in\mathcal{B}} \hat{u}_p^k \exp\left(2i\pi ap/N\right)$$

with the Fourier coefficients defined by

$$\hat{u}_p^k = \frac{1}{N}\sum_{b\in\mathcal{B}} U^k(x_b)\exp\left(-2i\pi bp/N\right).$$

We approximate the Laplacian by the Fourier differentiation matrix $D^{(2)}$ which for $j,l = 0,\dots,N-1$ takes the form

$$\left(D^{(2)}\right)_{jl} = \begin{cases} \frac{1}{4}(-1)^{j+1}N + \frac{(-1)^{j+l+1}}{2\sin^2\left(\frac{(j-l)\pi}{N}\right)}, & \text{if} \quad j\neq l \\ -\frac{(N-1)(N-2)}{12}, & \text{otherwise.} \end{cases}$$

However, to avoid the need to use the symmetric matrix $D^{(2)}$ in the previous form and gain in computational time, it can be preferable to use the method proposed in [18] that uses the fact that the previous differentiation matrix is diagonal in Fourier space. Therefore, inverting this matrix has a very low cost. In our work, since we use the Hamiltonian system (7.2) to analyze the properties of the SAV scheme we will use the previously defined differentiation matrix $D^{(2)}$.

For the N collocation points x_a, we define the interpolation operation I_N by

$$(I_N u)(x) = \sum_{p\in\mathcal{B}} \tilde{u}_p e^{2i\pi xp/N}.$$

We have the following interpolation error (see Section 5.8.1 in [50]):

Lemma 42 (Interpolation error) *For any $u \in C(0,T;H_p^m(\Omega))$ with $d \le 3$, we have*

$$\begin{cases} \|I_N u - u\|_{H_p^l(\Omega))} \le CN^{l-m}\left|u\right|_m, \quad 0\le l\le m, \\ \|I_N\partial_t u - \partial_t u\|_{H_p^l(\Omega))} \le CN^{l-m}\left|\partial_t u\right|_m, \quad 0\le l\le m. \end{cases}$$

7.2.2 The fully discrete SAV scheme

Applying the time discretization described in the previous section, for $k = 0 \to N_T$, the semi-discrete model of (7.4) reads

$$\begin{cases} \frac{p^{k+1}-p^k}{\tau} & = -\Delta q^{k+1/2} + r^{k+1/2}\tilde{g}_1^{k+1/2}, \\ \frac{q^{k+1}-q^k}{\tau} & = \Delta p^{k+1/2} - r^{k+1/2}\tilde{g}_2^{k+1/2}, \\ r^{k+1}-r^k & = \frac{1}{2}\left[\left(\tilde{g}_1^{k+1/2}, q^{k+1}-q^k\right) + \left(\tilde{g}_2^{k+1/2}, p^{k+1}-p^k\right)\right], \end{cases} \tag{7.6}$$

where $\phi^{k+1/2} = \left(\phi^{k+1}+\phi^k\right)/2$ and $\tilde{g}_i^{k+1/2}$ is a second order extrapolation of g_i at time $t = t^{k+1/2}$.

Denoting by capital letters the vectors of unknowns P^k, Q^k that are approximations at each collocation nodes of the continuous (in space) unknowns p^k, q^k, the fully discrete space-time

scheme then takes the form

$$\begin{cases} \frac{P^{k+1}-P^k}{\tau} &= -D^{(2)}Q^{k+1/2} + R^{k+1/2}\tilde{G}_1^{k+1/2}, \\ \frac{Q^{k+1}-Q^k}{\tau} &= D^{(2)}P^{k+1/2} - R^{k+1/2}\tilde{G}_2^{k+1/2}, \\ r^{k+1} - r^k &= \frac{1}{2}\left[\left(\tilde{G}_1^{k+1/2}, Q^{k+1}-Q^k\right) + \left(\tilde{G}_2^{k+1/2}, P^{k+1}-P^k\right)\right], \end{cases} \tag{7.7}$$

where $\tilde{G}_1, \tilde{G}_2$ are the vectors associated to the functions $\tilde{g}_1$ and $\tilde{g}_2$.

Let us now present two algorithms for the efficient solution of the fully discrete SAV system (7.7). The two methods are equivalent and reduce the problem to the solving of two linear systems that involves only real variables.

Algorithm 1

The algorithm below was originally proposed for solving the fully discrete SAV system arising in gradient flows [181].

Let us give the procedure on how to solve the system (7.7). First, we need to replace r^{k+1} in the first two equations using the third equation. This yields that

$$\begin{cases} \left(P^{k+1}-P^k\right) = -\tau D^{(2)}\dfrac{\left(Q^{k+1}+Q^k\right)}{2} \\ \qquad + \tau\left(r^k + \dfrac{1}{4}\left[\left(\tilde{G}_1^{k+1/2}, Q^{k+1}-Q^k\right) + \left(\tilde{G}_2^{k+1/2}, P^{k+1}-P^k\right)\right]\right)\tilde{G}_1^{k+1/2}, \\ \left(Q^{k+1}-Q^k\right) = \tau D^{(2)}\dfrac{\left(P^{k+1}-P^k\right)}{2} \\ \qquad - \tau\left(r^k + \dfrac{1}{4}\left[\left(\tilde{G}_1^{k+1/2}, Q^{k+1}-Q^k\right) + \left(\tilde{G}_2^{k+1/2}, P^{k+1}-P^k\right)\right]\right)\tilde{G}_2^{k+1/2}. \end{cases}$$

Next we set

$$Z^k = \begin{pmatrix} P^k \\ Q^k \end{pmatrix}, \tilde{G}^{k+1/2} = \begin{pmatrix} \tilde{G}_2^{k+1/2} \\ \tilde{G}_1^{k+1/2} \end{pmatrix}, \tilde{B}^{k+1/2} = \begin{pmatrix} -\tilde{G}_1^{k+1/2} \\ \tilde{G}_2^{k+1/2} \end{pmatrix}.$$

This allows us to rewrite the system into a matrix form

$$AZ^{k+1} + \frac{\tau}{4}\left(\tilde{G}^{k+1/2}, Z^{k+1}\right)\tilde{B}^{k+1/2} = C^k, \tag{7.8}$$

where

$$A = \begin{bmatrix} I & \frac{\tau}{2}D^{(2)} \\ -\frac{\tau}{2}D^{(2)} & I \end{bmatrix}$$

and

$$C^k = \begin{pmatrix} I & -\frac{\tau}{2}D^{(2)} \\ \frac{\tau}{2}D^{(2)} & I \end{pmatrix} Z^k - \tau r^k \tilde{B}^{k+1/2} + \frac{\tau}{4}\left(\tilde{G}^{k+1/2}, Z^k\right)\tilde{B}^{k+1/2},$$

with I the identity matrix. Multiplying (7.8) by A^{-1} and taking the discrete inner product with $\tilde{G}^{k+1/2}$, we finally obtain

$$\left(\tilde{G}^{k+1/2}, Z^{k+1}\right) = \frac{\left(\tilde{G}^{k+1/2}, A^{-1}C^k\right)}{1 + \frac{\tau}{4}\left(\tilde{G}^{k+1/2}, A^{-1}\tilde{B}^{k+1/2}\right)}. \tag{7.9}$$

Then, knowing $\left(\tilde{G}^{k+1/2}, Z^{k+1}\right)$, Z^{k+1} is computed using (7.8) and r^{k+1} is calculated from the third equation of (7.7). Therefore, solving the fully discrete SAV model (7.7) reduces to solving the linear system constituted by the equations (7.9) and (7.8).

Algorithm 2

Below we describe a second algorithm recently proposed in [18] for the numerical solution of the fully discrete NLS SAV scheme (7.7). Rewriting the scheme in its matrix form we have

$$\begin{cases} \frac{Z^{k+1}-Z^k}{\tau} & = -JZ^{k+1/2} - r^{k+1/2}\tilde{B}^{k+1/2}, \\ r^{k+1}-r^k & = \frac{1}{2}\left[\left(\tilde{G}_1^{k+1/2}, Q^{k+1}-Q^k\right) + \left(\tilde{G}_2^{k+1/2}, P^{k+1}-P^k\right)\right], \end{cases} \tag{7.10}$$

where

$$J = \begin{bmatrix} 0 & D^{(2)} \\ -D^{(2)} & 0 \end{bmatrix}.$$

Using the decomposition

$$Z^{k+1/2} = Z_1^{k+1/2} + r^{k+1/2} Z_2^{k+1/2}, \tag{7.11}$$

and adding $\frac{2}{\tau}Z^k$ on both sides of the first equation of (7.10), we furthermore obtain that

$$\frac{2}{\tau}\left[Z^{k+1/2} + r^{k+1/2}Z_2^{k+1/2}\right] = \frac{2}{\tau}Z^k - J\left[Z_1^{k+1/2} + r^{k+1/2}Z_2^{k+1/2}\right] - r^{k+1/2}\tilde{B}^{k+1/2}. \tag{7.12}$$

Applying the same decomposition to the second equation of (7.10) and adding $2r^k$ on both sides, we get

$$2r^{k+1/2} = 2r^k + \left(\tilde{G}^{k+1/2}, \left[Z_1^{k+1/2} + r^{k+1/2}Z_2^{k+1/2}\right] - Z^k\right). \tag{7.13}$$

Hence, denoting by I the identity matrix, we first solve the equation (7.12) using the system

$$\begin{cases} \left[\frac{2}{\tau}I + J\right] Z_1^{k+1/2} = \frac{2}{\tau}Z^k \\ \left[\frac{2}{\tau}I + J\right] Z_2^{k+1/2} = -\tilde{B}^{k+1/2}. \end{cases} \tag{7.14}$$

Then we compute $r^{k+1/2}$ by solving equation (7.13) which yields that

$$r^{k+1/2} = \frac{2r^k + \left(\tilde{G}^{k+1/2}, Z_1^{k+1/2} - Z^k\right)}{2 - \left(\tilde{G}^{k+1/2}, Z_2^{k+1/2}\right)}.$$

From the decomposition (7.11) we get $Z^{k+1/2}$ from which we compute Z^{k+1} and r^{k+1}.

Remark 43 *Since in practice the computation and storage of the invert of a non-diagonal matrix has to be avoided, Algorithm 2 is a preferable choice. Indeed, the main step in Algorithm 2 lies in solving two decoupled linear equations* (7.14)*. To do so, standard tools of linear systems can be applied such as matrix-free preconditioned Krylov solvers. We refer to Appendix C in [178] for a description of iterative solvers of linear system and preconditioning.*

Even though the previous remark already highlights the main advantage of Algorithm 2, we emphasize that the inversion of the main matrix in Algorithm 1 can be carried out efficiently.

Remark 44 *Referring to [181], we remark that the inversion of the matrix A in the first algorithm and the matrix $\left[\frac{2}{\tau} + J\right]$ in the second Algorithm can be carried out efficiently using the*

Sherman-Morrison-Woodbury formula [103]

$$(A+UV^T)^{-1} = A^{-1} - A^{-1}U(I+V^TA^{-1}U)^{-1}V^TA^{-1},$$

where A is a $n \times n$ and U, V are $n \times k$ matrices, and I is the $k \times k$ identity matrix.

Remark 45 *Referring to [93], a fast solver for solving the linear system (7.8) (in Algorithm 1) and (7.14) (in Algorithm 2) exists. It uses the fact that the differentiation matrix $D^{(2)}$ can be decomposed into $D^{(2)} = F^{-1}\Lambda F$ where F and F^{-1} are the corresponding matrices for the discrete Fourier transformation and Λ is a diagonal matrix with eigenvalues of $D^{(2)}$ as its entries. Therefore, the matrix A from Equation (7.8) admits the decomposition*

$$A = F^{-1}MF, \quad \text{with} \quad M = \begin{bmatrix} I & \frac{\tau}{2}\Lambda \\ -\frac{\tau}{2}\Lambda & I \end{bmatrix}.$$

We note that a similar decomposition exists for the matrix $[\frac{2}{\tau}I + J]$ in equation (7.14). Thanks to the above decomposition, the inverse of the matrix A can be computed explicitly in an efficient manner since

$$A^{-1} = F^{-1}M^{-1}F, \quad \text{and} \quad M^{-1} = M^T \begin{bmatrix} (I+\frac{\tau^2}{4}\Lambda^2)^{-1} & 0 \\ 0 & (I+\frac{\tau^2}{4}\Lambda^2)^{-1} \end{bmatrix},$$

where $(I + \frac{\tau^2}{4}\Lambda^2)$ is a diagonal matrix, such that its inverse is fast to compute.

7.3 Conservation properties and inequalities

In this section we outline the conserved quantities of the SAV method. It is well known that due to its design the SAV scheme preserves a modified version of the underlying Hamiltonian. In addition, to the conservation of energy, there is a wide variety of properties in the continuous equation which is feasible to preserve also on the numerical (discrete) level, we refer to Bao and Cai [22] as well as Antoine *et al.* [15]: *i) time-reversibility or symmetry, i.e.* the system is unchanged when $\tau \to -\tau$, *ii) gauge-invariance, i.e.* if the potential V is changed such that $V \to V + \alpha$ with α a real constant then the density $|u|^2$ remains unchanged, *iii)* conservation of mass, *i.e.* $\|u(t)\|_{L^2(\Omega)} = \|u(0)\|_{L^2(\Omega)}$, and the Hamiltonian energy, *i.e.* $H(t) = H(0)$, *iv)* preservation of the dispersion relation

$$\omega(k) = \frac{|k|^2}{2} + f(|A|^2),$$

for the plane wave solutions $u(t,x) = Ae^{ik\cdot x - \omega t}$.

Proving analytically that the SAV scheme for the NLS equation meets the points *ii)* and *iv)* (over long time scales) is up to our knowledge not possible with current techniques. However, the other points can be verified for a large number of nonlinearities. Here, we briefly recall the proofs of the conservation properties and refer to [93, 18], where they have been first set in context of nonlinear Schrödinger equations and our Theorems 46 and 47 are found by a combination of the results from [93] and [18].

Theorem 46 (Conservation of the modified discrete energy) *The scheme (7.6) is asso-*

ciated to the discrete modified Hamiltonian

$$\tilde{H}^{k+1} = \frac{1}{2}\left(\left|Q^{k+1}\right|_1^2 + \left|P^{k+1}\right|_1^2\right) + \left|r^{k+1}\right|^2, \tag{7.15}$$

and conserves the modified Hamiltonian energy through time i.e.

$$\tilde{H}^{k+1} = \tilde{H}^{k}. \tag{7.16}$$

Proof. Taking the inner product with $Q^{k+1} - Q^k$ for the first equation of (7.7) and for the second with $-\left(P^{k+1} - P^k\right)$, then summing the results we get

$$\begin{aligned} 0 = \frac{1}{2}&\left(\left|Q^{k+1}\right|_1^2 - \left|Q^{k}\right|_1^2 + \left|P^{k+1}\right|_1^2 - \left|P^{k}\right|_1^2\right) \\ &+ r^{k+1/2}\left[\left(\tilde{G}_1^{k+1/2}, Q^{k+1} - Q^k\right) + \left(\tilde{G}_2^{k+1/2}, P^{k+1} - P^k\right)\right], \end{aligned}$$

where $|\cdot|_1 = \|\nabla\cdot\|_0$ is the H^1-seminorm. Then, multiplying the third equation of (7.7) by $2R^{k+1/2}$ and using the result in the previous equation, we obtain

$$0 = \frac{1}{2}\left(\left|Q^{k+1}\right|_1^2 - \left|Q^{k}\right|_1^2 + \left|P^{k+1}\right|_1^2 - \left|P^{k}\right|_1^2\right) + \left(\left|r^{k+1}\right|^2 - \left|r^{k}\right|^2\right),$$

from which we can conclude both (7.15) and (7.16). □

The SAV scheme also preserves the mass up to an error of order $\mathcal{O}(\tau^3)$, where the latter error is introduced by the second-order extrapolation.

Theorem 47 (Conservation of the L^2 norm) *The scheme (7.4) conserves the L^2 norm of the solution up to an order $\mathcal{O}(\tau^3)$ i.e.*

$$\left\|U^{k+1}\right\|_0^2 = \left\|U^{k}\right\|_0^2 + \mathcal{O}(\tau^3), \tag{7.17}$$

with $U^k = P^k + iQ^k$.

Proof. Taking the inner product of first equation of (7.7) with $2P^{k+1/2}$, the second equation with $2Q^{k+1/2}$, and summing the two we get

$$\begin{aligned} &\frac{1}{\tau}\left(\left\|P^{k+1}\right\|_0^2 - \left\|P^{k}\right\|_0^2 + \left\|Q^{k+1}\right\|_0^2 - \left\|Q^{k}\right\|_0^2\right) \\ &\quad = 2r^{k+1/2}\left(-\left(\tilde{G}_2^{k+1/2}, Q^{k+1/2}\right) + \left(\tilde{G}_1^{k+1/2}, P^{k+1/2}\right)\right). \end{aligned}$$

Since $\tilde{G}_i^{k+1/2}$ is a second-order approximation of $G_i^{k+1/2}$, we can write

$$\frac{1}{\tau}\left(\left\|U^{k+1}\right\|_0^2 - \left\|U^{k}\right\|_0^2\right) = 2r^{k+1/2}\left(-\left(G_2^{k+1/2}, Q^{k+1/2}\right) + \left(G_1^{k+1/2}, P^{k+1/2}\right)\right) + \mathcal{O}(\tau^2).$$

Then, we find that

$$\begin{aligned} G_1^{k+1/2} &= \frac{1}{\sqrt{\mathcal{E}_1\left(P^{k+1/2}, Q^{k+1/2}\right) + \mathcal{E}_c}} \frac{\partial \mathcal{E}_1\left(P^{k+1/2}, Q^{k+1/2}\right)}{\partial Q^{k+1/2}} \\ &= V(x)Q^{k+1/2} + f\left(\left|P^{k+1/2}\right|^2, \left|Q^{k+1/2}\right|^2\right)Q^{k+1/2}, \end{aligned}$$

and

$$G_2^{k+1/2} = \frac{1}{\sqrt{\mathcal{E}_1\left(P^{k+1/2}, Q^{k+1/2}\right) + \mathcal{E}_c}} \frac{\partial \mathcal{E}_1\left(P^{k+1/2}, Q^{k+1/2}\right)}{\partial P^{k+1/2}}$$
$$= V(x)P^{k+1/2} + f\left(\left|P^{k+1/2}\right|^2, \left|Q^{k+1/2}\right|^2\right) P^{k+1/2},$$

from which we easily obtain

$$-\left(G_2^{k+1/2}, Q^{k+1/2}\right) + \left(G_1^{k+1/2}, P^{k+1/2}\right) = 0.$$

Consequently, we obtain (7.17). □

To derive H^2-bound for the solution of the SAV scheme, we use the following proposition. The proof of this technical result can be found in Lemma 2.3 in [179].

Proposition 48 (Bound for $\left\|\nabla G_i^{k+1/2}\right\|_0$) *Assume that the functions g_i (i=1,2) satisfy (7.5) and let $\|U\|_1 \leq M$ for some constant $M > 0$. Then there exists $0 \leq \sigma < 1$ such that*

$$\left\|\nabla G_i^{k+1/2}\right\| \leq C(M)\left(1 + \left\|\nabla \Delta P^{k+1/2}\right\|_0^{2\sigma} + \left\|\nabla \Delta Q^{k+1/2}\right\|_0^{2\sigma}\right). \tag{7.18}$$

We have the following result on the H^2-norm of $P^{k+1/2}$ and $Q^{k+1/2}$.

Proposition 49 (H^2-bound on the numerical solution) *The solution $\{P^{k+1}, Q^{k+1}\}$ of (7.6) satisfies*

$$\max_{k=1,\ldots,N_T-1} \left\|\Delta P^{k+1}\right\|_0^2 + \left\|\Delta Q^{k+1}\right\|_0^2 \leq CT + \left\|\Delta P^0\right\|_0^2 + \left\|\Delta Q^0\right\|_0^2. \tag{7.19}$$

Proof. First, we multiply the first equation of (7.6) by $\Delta^2(Q^{k+1/2})$, the second equation by $\Delta^2(P^{k+1/2})$ and integrate over Ω. Then, by summing the two, we obtain, after integration by parts, that

$$\left\|\nabla \Delta Q^{k+1/2}\right\|_0^2 + \left\|\nabla \Delta P^{k+1/2}\right\|_0^2 = \left(r^{k+1/2}\nabla \tilde{G}_1^{k+\frac{1}{2}}, \nabla \Delta Q^{k+1/2})\right) + \left(r^{k+1/2}\nabla \tilde{G}_2^{k+\frac{1}{2}}, \nabla \Delta P^{k+1/2}\right).$$

From the conservation of the modified Hamiltonian (7.15)–(7.16) and assuming a finite initial Hamiltonian, we have

$$\left(r^{k+1/2}\nabla \tilde{G}_1^{k+\frac{1}{2}}, \nabla \Delta Q^{k+1/2}\right) + \left(r^{k+1/2}\nabla \tilde{G}_2^{k+\frac{1}{2}}, \nabla \Delta P^{k+1/2}\right)$$
$$\leq \frac{C}{2}\left(\left\|\nabla \tilde{G}_1^{k+\frac{1}{2}}\right\|_0^2 + \left\|\nabla \Delta Q^{k+1/2}\right\|_0^2 + \left\|\nabla \tilde{G}_2^{k+\frac{1}{2}}\right\|_0^2 + \left\|\nabla \Delta P^{k+1/2}\right\|_0^2\right).$$

Then, from the result of Proposition 48, for any $\epsilon > 0$, we have

$$\left\|\nabla \tilde{G}_1^{k+\frac{1}{2}}\right\|_0^2 + \left\|\nabla \tilde{G}_2^{k+\frac{1}{2}}\right\|_0^2 \leq \epsilon \left\|\nabla \Delta Q^{k+1/2}\right\|_0^2 + \epsilon \left\|\nabla \Delta P^{k+1/2}\right\|_0^2 + C(\epsilon).$$

Therefore, combining the two previous inequalities, we obtain

$$\left\|\nabla \Delta Q^{k+1/2}\right\|_0^2 + \left\|\nabla \Delta P^{k+1/2}\right\|_0^2 \leq C. \tag{7.20}$$

Secondly, by multiplying the first equation of (7.6) with $\Delta^2(P^{k+1/2})$, the second equation with

$\Delta^2(Q^{k+1/2})$, integrating over Ω, and summing the two, we obtain after integration by parts that

$$\begin{aligned}\left\|\Delta P^{k+1}\right\|_0^2 - \left\|\Delta P^k\right\|_0^2 + \left\|\Delta Q^{k+1}\right\|_0^2 - \left\|\Delta Q^k\right\|_0^2 \\ = \tau r^{k+\frac{1}{2}} \left(\nabla \tilde{G}_1^{k+\frac{1}{2}}, \nabla \Delta Q^{k+1/2}\right) - \tau r^{k+\frac{1}{2}} \left(\nabla \tilde{G}_2^{k+\frac{1}{2}}, \nabla \Delta P^{k+1/2}\right).\end{aligned}$$

Then, combining the result of Proposition 48 and the inequality (7.20), we have

$$\left\|\Delta P^{k+1}\right\|_0^2 - \left\|\Delta P^k\right\|_0^2 + \left\|\Delta Q^{k+1}\right\|_0^2 - \left\|\Delta Q^k\right\|_0^2 \leq \tau C,$$

and summing from $k = 0 \to N_T$, we obtain (7.19). □

Remark 50 *From the fact that $H^2(\Omega) \subseteq L^\infty(\Omega)$ for $d \leq 3$, we can conclude from the previous proposition that for $k = 1, \dots N_T - 1$,*

$$\left\|P^{k+1}\right\|_{L^\infty} + \left\|Q^{k+1}\right\|_{L^\infty} \leq C. \tag{7.21}$$

Next, we present the stability inequality that will be useful in the convergence analysis.

Proposition 51 (Stability inequality) *The solution of (7.6) satisfies the stability inequality*

$$\begin{aligned}\max_{k=0,\dots,N_T-1} \left[\left\|P^{k+1}\right\|_0^2 + \left\|Q^{k+1}\right\|_0^2\right] + \tau^2 \sum_{k=0}^{N_T-1} \left[\left\|\frac{P^{k+1}-P^k}{\tau}\right\|_0^2 + \left\|\frac{Q^{k+1}-Q^k}{\tau}\right\|_0^2\right] \\ \leq C(\tau, H^0, N_T).\end{aligned} \tag{7.22}$$

Proof. Multiplying the first equation with $2\tau P^{k+1}$, integrating over Ω and using $2(a-b)a = a^2 - b^2 + (a-b)^2$, we obtain

$$\left\|P^{k+1}\right\|_0^2 + \tau^2 \left\|\frac{P^{k+1}-P^k}{\tau}\right\|_0^2 - \left\|P^k\right\|_0^2 = -2\tau \left(\nabla Q^{k+1/2}, \nabla P^{k+1}\right) + 2\tau r^{k+1/2} \left(\tilde{G}_1^{k+1/2}, P^{k+1}\right).$$

Using the Cauchy-Schwartz inequality and (7.15)–(7.16), we obtain

$$-2\tau \left(\nabla Q^{k+1/2}, \nabla P^{k+1}\right) \leq 2\tau \left\|\nabla Q^{k+1/2}\right\|_0 \left\|\nabla P^{k+1}\right\|_0 \leq 4\tau H^0.$$

Then, from the conservation of the Hamiltonian (7.15)–(7.16), and the conservation of the L^2-norm of the solution (7.17), we obtain using the Cauchy-Schwartz inequality

$$r^{k+1/2} \left(\tilde{G}_1^{k+1/2}, P^{k+1}\right) \leq C \left(\left\|\tilde{G}_1^{k+1/2}\right\|_0 \left\|P^{k+1}\right\|_0\right) \leq C \left\|G_1^{k+1/2}\right\|_0 + \mathcal{O}(\tau^2).$$

Since from Proposition 49 and (7.21), we have that $\left\|G_i^{k+1/2}\right\|_0 \leq C$ with $i = 1, 2$, for a large number of nonlinearities. Therefore, combining the previous inequalities for the right-hand side of (7.3), we obtain

$$\left\|P^{k+1}\right\|_0^2 + \tau^2 \left\|\frac{P^{k+1}-P^k}{\tau}\right\|_0^2 \leq C\tau + \left\|P^k\right\|_0^2.$$

The same can be found for the second equation by repeating the same calculations. Summing for $k = 0 \to N_T - 1$, we find (7.22). □

7.4 Convergence analysis

7.4.1 Notations

To study the convergence of the scheme, we introduce the following notation: For $k = 0, \dots, N_T - 1$ we set

$$U(t,x) := \frac{t - t^k}{\tau} U^{k+1} + \frac{t^{k+1} - t}{\tau} U^k, \quad t \in (t^k, t^{k+1}],$$

and

$$\frac{\partial U}{\partial t} := \frac{U^{k+1} - U^k}{\tau} \quad t \in (t^k, t^{k+1}].$$

We also define

$$U^+ := U^{k+1}, \quad U^- := U^k,$$

and

$$U - U^+ = (t - t^{k+1}) \frac{\partial U}{\partial t}, \quad U - U^- = (t - t^k) \frac{\partial U}{\partial t} \quad t \in (t^k, t^{k+1}], \quad k \geq 0.$$

In addition, we take analogous definitions for P and Q: For $k = 0, \dots, K_T - 1$ we set

$$P(t,x) := \frac{t - t^k}{\tau} P^{k+1} + \frac{t^{k+1} - t}{\tau} P^k, \quad t \in (t^k, t^{k+1}],$$

$$\frac{\partial P}{\partial t} := \frac{P^{k+1} - P^k}{\tau} \quad t \in (t^k, t^{k+1}],$$

$$P^+ := P^{k+1}, \quad P^- := P^k,$$

and

$$P - P^+ = (t - t^{k+1}) \frac{\partial P}{\partial t}, \quad \text{and} \quad P - P^- = (t - t^k) \frac{\partial P}{\partial t} \quad t \in (t^k, t^{k+1}], \quad k \geq 0.$$

7.4.2 Convergence theorem

Now we are in the position to establish time convergence for the semi discrete SAV scheme (7.6).

Theorem 52 (Convergence) *Let $\{p, q\}$ be a pair of functions such that*

$$\begin{cases} p(t,x) \in L^2\left([0,T]; H^1(\Omega)\right) \bigcap H^1\left([0,T]; \left(H^1(\Omega)\right)'\right) \\ q(t,x) \in L^2\left([0,T]; H^1(\Omega)\right) \bigcap H^1\left([0,T]; \left(H^1(\Omega)\right)'\right). \end{cases}$$

Then for $\tau \to 0$ we can extract a subsequence of solutions of (7.6), such that

$$P, P^{\pm} \to p \quad \textit{strongly in} \quad L^2\left([0,T]; L^2(\Omega)\right), \tag{7.23}$$
$$Q, Q^{\pm} \to q \quad \textit{strongly in} \quad L^2\left([0,T]; L^2(\Omega)\right), \tag{7.24}$$
$$P, P^{\pm} \rightharpoonup p \quad \textit{weakly in} \quad L^2\left([0,T]; H^1(\Omega)\right), \tag{7.25}$$
$$Q, Q^{\pm} \rightharpoonup q \quad \textit{weakly in} \quad L^2\left([0,T]; H^1(\Omega)\right), \tag{7.26}$$
$$\frac{\partial P}{\partial t} \rightharpoonup \frac{\partial p}{\partial t} \quad \textit{weakly in} \quad L^2\left([0,T]; \left(H^1(\Omega)\right)'\right), \tag{7.27}$$
$$\frac{\partial Q}{\partial t} \rightharpoonup \frac{\partial q}{\partial t} \quad \textit{weakly in} \quad L^2\left([0,T]; \left(H^1(\Omega)\right)'\right), \tag{7.28}$$
$$r^{k+1} \rightharpoonup r(t) = \sqrt{\mathcal{E}_1[t] + \mathcal{E}_c} \quad \textit{weak-star in} \quad L^{\infty}(0,T). \tag{7.29}$$

The limit $\{p, q\}$ satisfies the nonlinear Schrödinger model (7.2) in the following weak sense

$$\begin{cases} \int_0^T \left\langle \frac{\partial p}{\partial t}, \eta \right\rangle \mathrm{d}t &= \int_0^T \int_\Omega \nabla q \nabla \eta + \left(V(x) q + \frac{\partial F(|p|^2, |q|^2)}{\partial q} \right) \eta \, \mathrm{d}x \, \mathrm{d}t \\ \int_0^T \left\langle \frac{\partial q}{\partial t}, \eta \right\rangle \mathrm{d}t &= \int_0^T \int_\Omega -\nabla p \nabla \eta - \left(V(x) p + \frac{\partial F(|p|^2, |q|^2)}{\partial p} \right) \eta \, \mathrm{d}x \, \mathrm{d}t, \end{cases} \tag{7.30}$$

for all $\eta \in L^2\left([0,T]; H^1(\Omega)\right)$.

Proof.

Step 1: Weak and strong convergences. First, the weak convergences (7.25), (7.26), (7.27) and (7.28) follow from the assumption that the initial Hamiltonian energy is bounded and the stability inequality (7.22).

Then, the weak-star convergence (7.29) also holds true by the conservation of the modified Hamiltonian and the boundedness of the initial state.

From the compact embedding $H^1(\Omega) \subset L^2(\Omega) \equiv \left(L^2(\Omega)\right)'$, we can apply the Lions-Aubin Lemma [135] to find both convergences (7.23) and (7.24).

Step 2: Limit system. Let us work on the first equation of the discrete system. We use a test function $\eta \in L^2\left([0,T]; H^1(\Omega)\right)$ and analyze the convergence of the terms separately. First, from the weak convergence (7.26), we have

$$\int_0^T \int_\Omega \nabla \left(\frac{Q^+ + Q^-}{2} \right) \nabla \eta \, \mathrm{d}x \, \mathrm{d}t \to \int_0^T \int_\Omega \nabla q \nabla \eta \, \mathrm{d}x \, \mathrm{d}t.$$

Secondly, from the fact that $\tilde{G}_1^{k+1/2}$ is a second-order approximation of $G_1^{k+1/2}$, we have

$$\int_0^T \int_\Omega r^{k+1/2} \tilde{G}_1^{k+1/2} \eta \, \mathrm{d}x \, \mathrm{d}t = \int_0^T \int_\Omega r^{k+1/2} G_1^{k+1/2} \eta \, \mathrm{d}x \, \mathrm{d}t + \int_0^T \int_\Omega r^{k+1/2} \mathcal{O}(\tau^2) \eta \, \mathrm{d}x \, \mathrm{d}t.$$

From the inequality

$$\left| \mathcal{E}_1(U^{k+1/2}) - \mathcal{E}_1(u) \right| \leq C \left\| U^{k+1/2} \right\|^2_{L^1(\Omega)},$$

and the fact that

$$P^{\pm}, Q^{\pm} \rightharpoonup p, q \quad \text{weak-star in} \quad L^{\infty}\left(0, T; H^1(\Omega)\right),$$

which follows from the conservation of both the Hamiltonian energy and the L^2 norm, we have

$$\mathcal{E}_1(U^{k+1/2}) \rightharpoonup \mathcal{E}_1(u) \quad \text{weak-star in} \quad L^{\infty}(0,T).$$

The same holds true for $\frac{\delta \mathcal{E}_1^{k+1/2}}{\delta q}$ and $\frac{\delta \mathcal{E}_1^{k+1/2}}{\delta p}$ using similar arguments. Then, using also the strong convergences (7.23) and (7.24), together with the weak-star convergence (7.29), we obtain

$$\int_0^T \int_\Omega r^{k+1/2} G_1^{k+1/2} \eta \,\mathrm{d}x \,\mathrm{d}t + \int_0^T \int_\Omega r^{k+1/2} \mathcal{O}(\tau^2) \eta \,\mathrm{d}x \,\mathrm{d}t \to \int_0^T \int_\Omega r(t) g_1(t) \eta \,\mathrm{d}x \,\mathrm{d}t.$$

Finally, for any $\eta \in H^1([0,T]; H^1(\Omega))$, by integration by parts we have

$$\int_0^T \left(\frac{\partial P}{\partial t}, \eta\right)^h \mathrm{d}t = -\int_0^T \left(P, \frac{\partial \eta}{\partial t}\right)^h \mathrm{d}t + (P(T), \eta(T))^h - (P(0), \eta(0))^h .$$

Hence, from the regularity of η and the convergence (7.23) , we obtain

$$\int_0^T \left(P, \frac{\partial \eta}{\partial t}\right)^h \mathrm{d}t \to \int_0^T \left(p, \frac{\partial \eta}{\partial t}\right) \mathrm{d}t \quad \text{as} \quad \tau \to 0 \quad \text{and} \quad \forall \eta \in H^1([0,T]; H^1(\Omega)).$$

Gathering the previous convergences, we have

$$(p(T), \eta(T))^h - (p(0), \eta(0))^h - \int_0^T \left(p, \frac{\partial \eta}{\partial t}\right) \mathrm{d}t = \int_0^T \int_\Omega \nabla q \nabla \eta + \left(V(x) q + \frac{\partial F(|p|^2, |q|^2)}{\partial q}\right) \eta \,\mathrm{d}x \,\mathrm{d}t.$$

Since $\nabla q + \left(V(x) q + \frac{\partial F(|p|^2, |q|^2)}{\partial q}\right) \in L^2(\Omega)$ which follow from the conservation of the Hamiltonian energy, we know that $p \in H^1\left([0,T]; H^{-1}(\Omega)\right)$. Finally, we find the first equation of the limit system (7.30) and the same arguments can be applied to the second equation. This yields the result.

□

7.5 Error analysis

In this section we analyse the difference between the exact and modified Hamiltonian, and establish a bound on

$$\left| H[p(t^k), q(t^k)] - \tilde{H}[P^k, Q^k] \right|.$$

In addition we prove second-order convergence of the fully discrete SAV scheme (7.7) approximating the solution of the nonlinear Schrödinger equation (7.1). We introduce the following notation to study the error

$$e_u^k = \theta_u^k + \rho_u^k, \tag{7.31}$$

where

$$\theta_u^k = U^k - (I_N u)(t^k, x), \quad \rho_u^k = (I_N u)(t^k, x) - u(t^k, x).$$

For our convergence result we assume that the solution u of (7.1) is sufficiently smooth satisfying

$$\|\partial_{ttt} u\|_{L^\infty(0,T;H^1(\Omega))} + \|u\|_{L^\infty(0,T;H^2(\Omega))} \leq C. \tag{7.32}$$

We define the different truncation errors by

$$T_u^{k+\frac{1}{2}} = \frac{u^{k+1}-u^k}{\Delta t} - \partial_t u(t^{k+\frac{1}{2}}),$$
$$\overline{T}_u^{k+\frac{1}{2}} = u^{k+\frac{1}{2}} - u(t^{k+\frac{1}{2}}) = \frac{u^{k+1}+u^k}{2} - u(t^{k+\frac{1}{2}}).$$

We commence with two important lemma that will be useful in the global error analysis.

Lemma 53 (Boundedness of nonlinear functions) *If (p,q) is a solution of (7.4) satisfying (7.32), we have for $i=1,2$*

$$|g_i(p,q)|, \left|\frac{\partial g_i}{\partial p}\right|, \left|\frac{\partial g_i}{\partial q}\right|, \left|\frac{\partial^2 g_i}{\partial p \partial q}\right|, \left|\frac{\partial^2 g_i}{\partial p^2}\right|, \left|\frac{\partial^2 g_i}{\partial q^2}\right| \leq C.$$

Proof. This result is found by a combination of the fact that $u \in L^\infty(0,T;H^2(\Omega))$, Remark 7.21, and assumption (7.5).

Remark 54 *From Lemma 53, and the hypothesis (7.32), we know that*

$$|\partial_{ttt} r| \leq C\left(\|\partial_{ttt} p\|_0^2 + \|\partial_{ttt} q\|_0^2\right).$$

We have the following Lemma on the norm of the truncation errors (see Lemma 4.7 in [198] for example).

Lemma 55 (Truncation errors) *For $\alpha = -1, 0, 1, 2$, we have*

$$\left\|T_\psi^{k+\frac{1}{2}}\right\|_{H^\alpha(\Omega)}^2 \leq \tau^3 \int_{t^k}^{t^{k+1}} \|\partial_{ttt}\psi(s)\|_{H^\alpha(\Omega)}^2 \,\mathrm{d}s,$$
$$\left\|\overline{T}_\psi^{k+\frac{1}{2}}\right\|_{H^\alpha(\Omega)}^2 \leq \tau^3 \int_{t^k}^{t^{k+1}} \|\partial_{ttt}\psi(s)\|_{H^\alpha(\Omega)}^2 \,\mathrm{d}s,$$

Theorem 56 (Error analysis) *Assume that the solution of (7.2) satisfies (7.32) with initial condition $u^0 \in H^3(\Omega)$. Then the discrete solution $\{P^{k+1}, Q^{k+1}\}$ of the fully discrete SAV scheme (7.7) satisfies the error estimate*

$$\frac{1}{2}\left\|\nabla e_q^{k+1}\right\|_0^2 + \frac{1}{2}\left\|\nabla e_p^{k+1}\right\|_0^2 + \left|e_r^{k+1}\right|^2 \leq C\exp\left(\left[1-C\tau\right]^{-1} t^{k+1}\right)\left(\tau^4 + N^{-4}\right),$$

where the constant C depends on the smoothness of the solution (7.32).

Proof.

Step 1: Error equations. We begin by evaluating the model (7.4) at time $t^{k+1/2}$

$$\begin{cases} \partial_t p(t^{k+1/2}) &= -\Delta q(t^{k+1/2}) + r(t^{k+1/2}) g_1(t^{k+1/2}), \\ \partial_t q(t^{k+1/2}) &= \Delta p(t^{k+1/2}) - r(t^{k+1/2}) g_2(t^{k+1/2}), \\ \frac{\mathrm{d}r}{\mathrm{d}t}(t^{k+1/2}) &= \frac{1}{2}\left[\left(g_1(t^{k+1/2}), \partial_t q(t^{k+1/2})\right) + \left(g_2(t^{k+1/2}), \partial_t p(t^{k+1/2})\right)\right]. \end{cases}$$

Subtracting the above equations from (7.7) yields

$$\begin{cases} \frac{e_p^{k+1}-e_p^k}{\tau} + T_p^{k+1/2} &= -\Delta\left(e_q^{k+1/2} + \overline{T}_q^{k+1/2}\right) + R^{k+1/2}\tilde{G}_1^{k+1/2} - r(t^{k+1/2})g_1(t^{k+1/2}), \\ \frac{e_q^{k+1}-e_q^k}{\tau} + T_q^{k+1/2} &= \Delta\left(e_p^{k+1/2} + \overline{T}_p^{k+1/2}\right) - R^{k+1/2}\tilde{G}_2^{k+1/2} + r(t^{k+1/2})g_2(t^{k+1/2}), \\ \frac{e_r^{k+1}-e_r^k}{\tau} + T_r^{k+1/2} &= \frac{1}{2}\big[\left(\tilde{G}_1^{k+1/2}, \frac{Q^{k+1}-Q^k}{\tau}\right) + \left(\tilde{G}_2^{k+1/2}, \frac{P^{k+1}-P^k}{\tau}\right) \\ & \quad - \left(g_1(t^{k+1/2}), \partial_t q(t^{k+1/2})\right) - \left(g_2(t^{k+1/2}), \partial_t p(t^{k+1/2})\right)\big]. \end{cases} \tag{7.33}$$

We introduce the error

$$e_{g,1}^{k+1/2} = \tilde{G}_1^{k+1/2} - g_1(t^{k+1/2}).$$

The rightmost terms of the two first equations of (7.33) can be replaced by

$$R^{k+1/2}\tilde{G}_1^{k+1/2} - r(t^{k+1/2})g_1(t^{k+1/2}) = \tilde{G}_1^{k+1/2}\left(e_r^{k+1/2} + \overline{T}_r^{k+1/2}\right) + r(t^{k+1/2})e_{g,1}^{k+1/2}, \tag{7.34}$$

and

$$R^{k+1/2}\tilde{G}_2^{k+1/2} - r(t^{k+1/2})g_2(t^{k+1/2}) = \tilde{G}_2^{k+1/2}\left(e_r^{k+1/2} + \overline{T}_r^{k+1/2}\right) + r(t^{k+1/2})e_{g,2}^{k+1/2}. \tag{7.35}$$

Similarly, we have

$$\begin{aligned} &\frac{1}{2}\left[\left(\tilde{G}_1^{k+1/2}, \frac{Q^{k+1}-Q^k}{\tau}\right) - \left(g_1(t^{k+1/2}), \partial_t q(t^{k+1/2})\right)\right] \\ &\quad = \frac{1}{2}\left[\left(\tilde{G}_1^{k+1/2}, \frac{e_q^{k+1}-e_q^k}{\tau} + T_q^{k+1/2}\right) + \left(e_{g,1}^{k+1/2}, \partial_t q(t^{k+1/2})\right)\right], \end{aligned} \tag{7.36}$$

and

$$\begin{aligned} &\frac{1}{2}\left[\left(\tilde{G}_2^{k+1/2}, \frac{P^{k+1}-P^k}{\tau}\right) - \left(g_2(t^{k+1/2}), \partial_t p(t^{k+1/2})\right)\right] \\ &\quad = \frac{1}{2}\left[\left(\tilde{G}_2^{k+1/2}, \frac{e_p^{k+1}-e_p^k}{\tau} + T_p^{k+1/2}\right) + \left(e_{g,2}^{k+1/2}, \partial_t p(t^{k+1/2})\right)\right]. \end{aligned} \tag{7.37}$$

Plugging (7.34), (7.35),(7.36), and (7.37) into (7.33), we thus obtain

$$\begin{cases} \frac{e_p^{k+1}-e_p^k}{\tau} + T_p^{k+1/2} &= -\Delta\left(e_q^{k+1/2} + \overline{T}_q^{k+1/2}\right) + \tilde{G}_1^{k+1/2}\left(e_r^{k+1/2} + \overline{T}_r^{k+1/2}\right) + r(t^{k+1/2})e_{g,1}^{k+1/2}, \\ \frac{e_q^{k+1}-e_q^k}{\tau} + T_q^{k+1/2} &= \Delta\left(e_p^{k+1/2} + \overline{T}_p^{k+1/2}\right) - \tilde{G}_2^{k+1/2}\left(e_r^{k+1/2} + \overline{T}_r^{k+1/2}\right) - r(t^{k+1/2})e_{g,2}^{k+1/2}, \\ \frac{e_r^{k+1}-e_r^k}{\tau} + T_r^{k+1/2} &= \frac{1}{2}\big[\left(\tilde{G}_1^{k+1/2}, \frac{e_q^{k+1}-e_q^k}{\tau} + T_q^{k+1/2}\right) + \left(e_{g,1}^{k+1/2}, \partial_t q(t^{k+1/2})\right) \\ & \quad + \left(\tilde{G}_2^{k+1/2}, \frac{e_p^{k+1}-e_p^k}{\tau} + T_p^{k+1/2}\right) + \left(e_{g,2}^{k+1/2}, \partial_t p(t^{k+1/2})\right)\big]. \end{cases}$$

Using the decomposition of the error (7.31), we furthermore obtain

$$\begin{cases} \frac{\theta_p^{k+1}-\theta_p^k}{\tau} + \Delta\left(\frac{\theta_q^{k+1}+\theta_q^k}{2}\right) & = -\frac{\rho_p^{k+1}-\rho_p^k}{\tau} - \Delta\left(\rho_q^{k+1/2} + \overline{T}_q^{k+1/2}\right) + \tilde{G}_1^{k+1/2}\left(e_r^{k+1/2} + \overline{T}_r^{k+1/2}\right) \\ & \quad + r(t^{k+1/2})e_{g,1}^{k+1/2} - T_p^{k+1/2}, \\ \frac{\theta_q^{k+1}-\theta_q^k}{\tau} - \Delta\left(\frac{\theta_p^{k+1}+\theta_p^k}{2}\right) & = -\frac{\rho_q^{k+1}-\rho_q^k}{\tau} + \Delta\left(\rho_p^{k+1/2} + \overline{T}_p^{k+1/2}\right) - \tilde{G}_2^{k+1/2}\left(e_r^{k+1/2} + \overline{T}_r^{k+1/2}\right) \\ & \quad - r(t^{k+1/2})e_{g,2}^{k+1/2} - T_q^{k+1/2}, \\ \frac{e_r^{k+1}-e_r^k}{\tau} + T_r^{k+1/2} & = \frac{1}{2}\Big[\left(\tilde{G}_1^{k+1/2}, \frac{e_q^{k+1}-e_q^k}{\tau} + T_q^{k+1/2}\right) + \left(e_{g,1}^{k+1/2}, \partial_t q(t^{k+1/2})\right) \\ & \quad + \left(\tilde{G}_2^{k+1/2}, \frac{e_p^{k+1}-e_p^k}{\tau} + T_p^{k+1/2}\right) + \left(e_{g,2}^{k+1/2}, \partial_t p(t^{k+1/2})\right)\Big]. \end{cases} \tag{7.38}$$

Step 2. Error estimate formula. We use the following notations to make the results more compact

$$D_\tau^1 \theta_p^{k+1} = \frac{\theta_p^{k+1} - \theta_p^k}{\tau}, \quad D^1 \theta_p^{k+1} = \theta_p^{k+1} - \theta_p^k.$$

Taking the inner product of the first equation of the system (7.38) with $-D^1\theta_q^{k+1}$ and the second with $D^1\theta_p^{k+1}$, and summing the results, we also have,

$$\begin{aligned} \frac{1}{2}D^1\left\|\nabla\theta_q^{k+1}\right\|_0^2 + \frac{1}{2}D^1\left\|\nabla\theta_p^{k+1}\right\|_0^2 &= \left(D_\tau^1\rho_p^{k+1}, D^1\theta_q^{k+1}\right) - \left(D_\tau^1\rho_q^{k+1}, D^1\theta_p^{k+1}\right) \\ &\quad - \left(\nabla\rho_q^{k+1/2}, \nabla D^1\theta_q^{k+1}\right) - \left(\nabla\rho_p^{k+1/2}, \nabla D^1\theta_p^{k+1}\right) \\ &\quad - \left(\nabla\overline{T}_q^{k+1/2}, \nabla D^1\theta_q^{k+1}\right) - \left(\nabla\overline{T}_p^{K+1/2}, \nabla D^1\theta_p^{k+1}\right) \\ &\quad - \left(\tilde{G}_1^{k+1/2}\left(e_r^{k+1/2} + \overline{T}_r^{k+1/2}\right), D^1\theta_q^{k+1}\right) \\ &\quad - \left(\tilde{G}_2^{k+1/2}\left(e_r^{k+1/2} + \overline{T}_r^{k+1/2}\right), D^1\theta_p^{k+1}\right) \\ &\quad - \left(r(t^{k+1/2})e_{g,1}^{k+1/2} - T_p^{k+1/2}, D^1\theta_q^{k+1}\right) \\ &\quad - \left(r(t^{k+1/2})e_{g,2}^{k+1/2} + T_p^{k+1/2}, D^1\theta_p^{k+1}\right). \end{aligned} \tag{7.39}$$

Multiplying the third equation of (7.38) by $2\tau e_r^{k+1/2}$, we have

$$\begin{aligned} &D^1\left|e_r^{k+1}\right|^2 + 2\tau T_r^{k+1/2}e_r^{k+1/2} - \tau e_r^{k+1/2}\Big[\left(\tilde{G}_1^{k+1/2}, D_\tau^1\rho_q^{k+1} + T_q^{k+1/2}\right) \\ &\quad + \left(\tilde{G}_2^{k+1/2}, D_\tau^1\rho_p^{k+1} + T_p^{k+1/2}\right) + \left(e_{g,1}^{k+1/2}, \partial_t q(t^{k+1/2})\right) + \left(e_{g,2}^{k+1/2}, \partial_t p(t^{k+1/2})\right)\Big] \\ &\quad = e_r^{k+1/2}\left[\left(\tilde{G}_1^{k+1/2}, D^1\theta_q^{k+1}\right) + \left(\tilde{G}_2^{k+1/2}, D^1\theta_p^{k+1}\right)\right]. \end{aligned} \tag{7.40}$$

Using (7.40) in (7.39), we have

$$\begin{aligned}
&\frac{1}{2}D^1\left\|\nabla\theta_q^{k+1}\right\|_0^2+\frac{1}{2}D^1\left\|\nabla\theta_p^{k+1}\right\|_0^2+D^1\left|e_r^{k+1}\right|^2\\
&\quad=\left(D_\tau^1\rho_p^{k+1},D^1\theta_q^{k+1}\right)-\left(D_\tau^1\rho_q^{k+1},D^1\theta_p^{k+1}\right)\\
&\quad-\left(\nabla\left(\rho_q^{k+1/2}+\overline{T}_q^{k+1/2}\right),\nabla D^1\theta_q^{k+1}\right)-\left(\nabla\left(\rho_p^{k+1/2}+\overline{T}_p^{k+1/2}\right),\nabla D^1\theta_p^{k+1}\right)\\
&\quad-\overline{T}_r^{k+1/2}\left[\left(\tilde{G}_1^{k+1/2},D^1\theta_q^{k+1}\right)+\left(\tilde{G}_2^{k+1/2},D^1\theta_p^{k+1}\right)\right]\\
&\quad-2\tau T_r^{k+1/2}e_r^{k+1/2}+\tau e_r^{k+1/2}\Big[\left(\tilde{G}_1^{k+1/2},D_\tau^1\rho_q^{k+1}+T_q^{k+1/2}\right)\\
&\quad+\left(\tilde{G}_2^{k+1/2},D_\tau^1\rho_p^{k+1}+T_p^{k+1/2}\right)+\left(e_{g,1}^{k+1/2},\partial_t q(t^{k+1/2})\right)+\left(e_{g,2}^{k+1/2},\partial_t p(t^{k+1/2})\right)\Big]\\
&\quad-\left(r(t^{k+1/2})e_{g,1}^{k+1/2}-T_p^{k+1/2},D^1\theta_q^{k+1}\right)\\
&\quad-\left(r(t^{k+1/2})e_{g,2}^{k+1/2}+T_p^{k+1/2},D^1\theta_p^{k+1}\right).
\end{aligned}\tag{7.41}$$

Step 3. Inequalities for the terms on the right-hand side of (7.41).

Now, we bound the right-hand side of (7.41). Using Lemma 42, Lemma 55 and Young's inequality we have

$$\left(D_\tau^1\rho_p^{k+1},D^1\theta_q^{k+1}\right)\le 4\left\|D_\tau^1\rho_p^{k+1}\right\|_0^2+\frac{1}{16}\left\|D^1\theta_q^{k+1}\right\|_0^2\le CN^{-6}+\frac{1}{16}\left\|D^1\theta_q^{k+1}\right\|_0^2,$$

$$-\left(D_\tau^1\rho_q^{k+1},D^1\theta_p^{k+1}\right)\le 4\left\|D_\tau^1\rho_q^{k+1}\right\|_0^2+\frac{1}{16}\left\|D^1\theta_p^{k+1}\right\|_0^2\le CN^{-6}+\frac{1}{16}\left\|D^1\theta_p^{k+1}\right\|_0^2.$$

Then, from Theorem (46), we have

$$\begin{aligned}
&-\left(\nabla\rho_q^{k+1/2},\nabla D^1\theta_q^{k+1}\right)-\left(\nabla\rho_p^{k+1/2},\nabla D^1\theta_p^{k+1}\right)\\
&\quad\le\left(\left\|\nabla\rho_q^{k+1/2}\right\|_0^2\left\|\nabla D^1\theta_q^{k+1}\right\|_0^2+\left\|\nabla\rho_p^{k+1/2}\right\|_0^2\left\|\nabla D^1\theta_p^{k+1}\right\|_0^2\right)\\
&\quad\le CN^{-4},
\end{aligned}$$

and

$$\begin{aligned}
&-\left(\nabla\overline{T}_q^{k+1/2},\nabla D^1\theta_q^{k+1}\right)-\left(\nabla\overline{T}_p^{K+1/2},\nabla D^1\theta_p^{k+1}\right)\\
&\quad\le\left(\left\|\nabla\overline{T}_q^{k+1/2}\right\|_0^2\left\|\nabla D^1\theta_q^{k+1}\right\|_0^2+\left\|\nabla\overline{T}_p^{K+1/2}\right\|_0^2\left\|\nabla D^1\theta_p^{k+1}\right\|_0^2\right)\\
&\quad\le C\tau^4,
\end{aligned}$$

For the rest of the terms on the right-hand side of (7.41), we use Lemma 55, and Proposition 51 together with Lemma 53, and Remark 54, to obtain

$$\begin{aligned}
&-\overline{T}_r^{k+1/2}\left[\left(\tilde{G}_1^{k+1/2},D^1\theta_q^{k+1}\right)+\left(\tilde{G}_2^{k+1/2},D^1\theta_p^{k+1}\right)\right]\\
&\quad\le 4\left|\overline{T}_r^{k+1/2}\right|^2\left(\left\|\tilde{G}_1^{k+1/2}\right\|_0^2+\left\|\tilde{G}_1^{k+1/2}\right\|_0^2\right)+\frac{1}{16}\left(\left\|D^1\theta_q^{k+1}\right\|_0^2+\left\|D^1\theta_p^{k+1}\right\|_0^2\right)\\
&\quad\le C\tau^4+\frac{1}{16}\left(\left\|D^1\theta_q^{k+1}\right\|_0^2+\left\|D^1\theta_p^{k+1}\right\|_0^2\right),
\end{aligned}\tag{7.42}$$

$$-2\tau T_r^{k+1/2} e_r^{k+1/2} \le C\tau \left(\left\| T_r^{k+1/2} \right\|_0^2 + \left| e_r^{k+1} \right|^2 + \left| e_r^k \right|^2 \right) \le C\tau^5 + \tau \left| e_r^{k+1} \right|^2 + \tau \left| e_r^k \right|^2,$$

$$\begin{aligned}
&\tau e_r^{k+1/2} \left(\tilde{G}_1^{k+1/2}, D_\tau^1 \rho_q^{k+1} + T_q^{k+1/2} \right) \\
&\quad \le \frac{\tau}{2} \left\| \tilde{G}_1^{k+1/2} \right\|_0^2 \left(\left\| D_\tau^1 \rho_q^{k+1} \right\|_0^2 + \left\| T_q^{k+1/2} \right\|_0^2 + \left| e_r^{k+1} \right|^2 + \left| e_r^k \right|^2 \right) \\
&\quad \le C\tau \left(N^{-6} + \tau^4 + \left| e_r^{k+1} \right|^2 + \left| e_r^k \right|^2 \right), \\
&\tau e_r^{k+1/2} \left(\tilde{G}_2^{k+1/2}, D_\tau^1 \rho_p^{k+1} + T_p^{k+1/2} \right) \\
&\quad \le \frac{\tau}{2} \left\| \tilde{G}_2^{k+1/2} \right\|_0^2 \left(\left\| D_\tau^1 \rho_p^{k+1} \right\|_0^2 + \left\| T_p^{k+1/2} \right\|_0^2 + \left| e_r^{k+1} \right|^2 + \left| e_r^k \right|^2 \right) \\
&\quad \le C\tau \left(N^{-6} + \tau^4 + \left| e_r^{k+1} \right|^2 + \left| e_r^k \right|^2 \right),
\end{aligned}$$

$$\begin{aligned}
&\tau e_r^{k+1/2} \left[\left(e_{g,1}^{k+1/2}, \partial_t q(t^{k+1/2}) \right) + \left(e_{g,2}^{k+1/2}, \partial_t p(t^{k+1/2}) \right) \right] \\
&\quad \le \frac{\tau}{2} \left\| \partial_t q(t^{k+1/2}) \right\|_0^2 \left(\left\| e_{g,1}^{k+1/2} \right\|_0^2 + \left| e_r^{k+1} \right|^2 + \left| e_r^k \right|^2 \right) \\
&\quad + \frac{\tau}{2} \left\| \partial_t p(t^{k+1/2}) \right\|_0^2 \left(\left\| e_{g,2}^{k+1/2} \right\|_0^2 + \left| e_r^{k+1} \right|^2 + \left| e_r^k \right|^2 \right) \\
&\quad \le C\tau \left(\left\| e_{g,1}^{k+1/2} \right\|_0^2 + \left\| e_{g,2}^{k+1/2} \right\|_0^2 + \left| e_r^{k+1} \right|^2 + \left| e_r^k \right|^2 \right),
\end{aligned}$$

and

$$\begin{aligned}
&- \left(r(t^{k+1/2}) e_{g,1}^{k+1/2} - T_p^{k+1/2}, D^1 \theta_q^{k+1} \right) - \left(r(t^{k+1/2}) e_{g,2}^{k+1/2} + T_p^{k+1/2}, D^1 \theta_p^{k+1} \right) \\
&\le 4 \left| r(t^{k+1/2} \right|^2 \left(\left\| e_{g,1}^{k+1/2} \right\|_0^2 + \left\| e_{g,2}^{k+1/2} \right\|_0^2 + \left\| T_p^{k+1/2} \right\|_0^2 + \left\| T_q^{k+1/2} \right\|_0^2 \right) \\
&+ \frac{1}{16} \left(\left\| D^1 \theta_q^{k+1} \right\|_0^2 + \left\| D^1 \theta_p^{k+1} \right\|_0^2 \right) \\
&\le C \left(\left\| e_{g,1}^{k+1/2} \right\|_0^2 + \left\| e_{g,2}^{k+1/2} \right\|_0^2 + 2\tau^4 \right) + \frac{1}{16} \left(\left\| D^1 \theta_q^{k+1} \right\|_0^2 + \left\| D^1 \theta_p^{k+1} \right\|_0^2 \right).
\end{aligned} \tag{7.43}$$

Step 4. Estimating the terms in the inequalities (7.42)–(7.43). First, we aim to emiminate the terms $\left\| D^1 \theta_p^{k+1} \right\|_0^2$ and $\left\| D^1 \theta_q^{k+1} \right\|_0^2$ in the above inequalities. Taking the inner product of the first equation of (7.38) with $2\tau\theta_p^{k+1}$, we obtain

$$\begin{aligned}
\left(D_\tau^1 \theta_p^{k+1}, 2\tau\theta_p^{k+1} \right) &= 2\tau \left(\nabla \theta_q^{k+1/2}, \nabla \theta_p^{k+1} \right) - 2\tau \left(D_\tau^1 \rho_p^{k+1}, \theta_p^{k+1} \right) + 2\tau \left(\nabla \left(\rho_q^{k+1/2} + \overline{T}_q^{k+1/2} \right), \nabla \theta_p^{k+1} \right) \\
&\quad + 2\tau \left(e_r^{k+1/2} + \overline{T}_r^{k+1/2} \right) \left(\tilde{G}_1^{k+1/2}, \theta_p^{k+1} \right) + 2\tau \left(r(t^{k+1/2}) e_{g,1}^{k+1/2} - T_p^{k+1/2}, \theta_p^{k+1} \right).
\end{aligned}$$

Knowing that

$$\left(D_\tau^1 \theta_p^{k+1}, 2\tau\theta_p^{k+1} \right) \ge \left\| D^1 \theta_p^{k+1} \right\|_0^2,$$

we have

$$\left\|D^1\theta_p^{k+1}\right\|_0^2 \leq 2\tau\left(\nabla\theta_q^{k+1/2},\nabla\theta_p^{k+1}\right) - 2\tau\left(D_\tau^1\rho_p^{k+1},\theta_p^{k+1}\right) + 2\tau\left(\nabla\left(\rho_q^{k+1/2}+\overline{T}_q^{k+1/2}\right),\nabla\theta_p^{k+1}\right) + 2\tau\left(e_r^{k+1/2}+\overline{T}_r^{k+1/2}\right)\left(\tilde{G}_1^{k+1/2},\theta_p^{k+1}\right) + 2\tau\left(r(t^{k+1/2})e_{g,1}^{k+1/2}-T_p^{k+1/2},\theta_p^{k+1}\right). \tag{7.44}$$

Let us bound the terms on the right-hand side of (7.44). Using Lemma 42 we find that

$$\begin{aligned}
2\tau\left(\nabla\theta_q^{k+1/2},\nabla\theta_p^{k+1}\right) &\leq \tau\left(\left\|\nabla\theta_q^{k+1}\right\|_0^2+\left\|\nabla\theta_q^{k}\right\|_0^2+\left\|\nabla\theta_p^{k+1}\right\|_0^2\right),\\
-2\tau\left(D_\tau^1\rho_p^{k+1},\theta_p^{k+1}\right) &\leq \tau\left(\left\|D_\tau^1\rho_p^{k+1}\right\|_{H^{-1}(\Omega)}^2+\left\|\nabla\theta_p^{k+1}\right\|_0^2\right) \leq \tau\left(N^{-6}+\left\|\nabla\theta_p^{k+1}\right\|_0^2\right),\\
2\tau\left(\nabla\left(\rho_q^{k+1/2}+\overline{T}_q^{k+1/2}\right),\nabla\theta_p^{k+1}\right) &\leq \tau\left(CN^{-4}+C\tau^4+\left\|\nabla\theta_p^{k+1}\right\|_0^2\right),\\
2\tau\left(r(t^{k+1/2})e_{g,1}^{k+1/2}-T_p^{k+1/2},\theta_p^{k+1}\right) &\leq \tau\left(C\left\|e_{g,1}^{k+1/2}\right\|_0^2+\tau^4+C\left\|\nabla\theta_p^{k+1}\right\|_0^2\right),
\end{aligned}$$

where we have used the Poincaré inequality to obtain the last inequality. Plugging the previous inequalities into (7.44), we obtain

$$\left\|D^1\theta_p^{k+1}\right\|_0^2 \leq \tau\left(\left\|\nabla\theta_q^{k+1}\right\|_0^2+\left\|\nabla\theta_q^{k}\right\|_0^2+CN^{-4}+C\tau^4+C\left\|e_{g,1}^{k+1/2}\right\|_0^2+C\left\|\nabla\theta_p^{k+1}\right\|_0^2\right). \tag{7.45}$$

Similarly, taking the inner product of the second equation of (7.38) with $2\tau\theta_q^{k+1}$ and repeating the same steps as before, we obtain

$$\left\|D^1\theta_q^{k+1}\right\|_0^2 \leq \tau\left(\left\|\nabla\theta_p^{k+1}\right\|_0^2+\left\|\nabla\theta_p^{k}\right\|_0^2+CN^{-4}+C\tau^4+C\left\|e_{g,2}^{k+1/2}\right\|_0^2+C\left\|\nabla\theta_q^{k+1}\right\|_0^2\right). \tag{7.46}$$

Step 5. Estimating $\left\|e_{g,1}^{k+1/2}\right\|_0^2$ *and* $\left\|e_{g,2}^{k+1/2}\right\|_0^2$. Using the notations

$$S(p,q)=\sqrt{\mathcal{E}_1(p,q)+C},$$

and

$$N_1(p,q)=\frac{\delta}{\delta q}\mathcal{E}_1(p,q),\quad N_2(p,q)=\frac{\delta}{\delta p}\mathcal{E}_1(p,q)$$

we have that

$$
\begin{aligned}
e_{g,1}^{k+1/2} &= G_1(P^{k+1/2}, Q^{k+1/2}) - g_1(p(t^{k+1/2}), q(t^{k+1/2})) \\
&= \frac{N_1(P^{k+1/2}, Q^{k+1/2})}{S(P^{k+1/2}, Q^{k+1/2})} - \frac{N_1(p(t^{k+1/2}), q(t^{k+1/2}))}{S(p(t^{k+1/2}), q(t^{k+1/2}))} \\
&= \frac{N_1(P^{k+1/2}, Q^{k+1/2})}{S(P^{k+1/2}, Q^{k+1/2})} - \frac{N_1(P^{k+1/2}, Q^{k+1/2})}{S(p(t^{k+1/2}), q(t^{k+1/2}))} + \frac{N_1(P^{k+1/2}, Q^{k+1/2})}{S(p(t^{k+1/2}), q(t^{k+1/2}))} \\
&\quad - \frac{N_1(p(t^{k+1/2}), q(t^{k+1/2}))}{S(p(t^{k+1/2}), q(t^{k+1/2}))} \\
&= \frac{N_1(P^{k+1/2}, Q^{k+1/2}) \left[\mathcal{E}_1(p(t^{k+1/2}), q(t^{k+1/2})) - \mathcal{E}_1(P^{k+1/2}, Q^{k+1/2})\right]}{S(P^{k+1/2}, Q^{k+1/2}) S(p(t^{k+1/2}), q(t^{k+1/2})) \left[S(P^{k+1/2}, Q^{k+1/2}) + S(p(t^{k+1/2}), q(t^{k+1/2}))\right]} \\
&\quad + \frac{N_1(P^{k+1/2}, Q^{k+1/2}) - N_1(p(t^{k+1/2}), q(t^{k+1/2}))}{S(p(t^{k+1/2}), q(t^{k+1/2}))}.
\end{aligned}
$$

From the smoothness assumption (7.32), Lemma 53, and Remark 50, we have

$$
\left\| e_{g,1}^{k+1/2} \right\|_0^2 \leq C \left[\left\| P^{k+1/2} - p(t^{k+1/2}) \right\|_0^2 + \left\| Q^{k+1/2} - q(t^{k+1/2}) \right\|_0^2 \right].
$$

Then, using the notation (7.31) and Lemma 55, we obtain

$$
\begin{aligned}
\left\| e_{g,1}^{k+1/2} \right\|_0^2 &\leq C \left[\left\| \theta_p^{k+1} \right\|_0^2 + \left\| \theta_p^k \right\|_0^2 + \left\| \theta_q^{k+1} \right\|_0^2 + \left\| \theta_q^k \right\|_0^2 + \tau^3 + N^{-4} \right] \\
&\leq C \left[\left\| \nabla\theta_p^{k+1} \right\|_0^2 + \left\| \nabla\theta_p^k \right\|_0^2 + \left\| \nabla\theta_q^{k+1} \right\|_0^2 + \left\| \nabla\theta_q^k \right\|_0^2 + \tau^3 + N^{-4} \right].
\end{aligned}
$$

Similarly, we have

$$
\left\| e_{g,2}^{k+1/2} \right\|_0^2 \leq C \left[\left\| \nabla\theta_q^{k+1} \right\|_0^2 + \left\| \nabla\theta_q^k \right\|_0^2 + \left\| \nabla\theta_p^{k+1} \right\|_0^2 + \left\| \nabla\theta_p^k \right\|_0^2 + \tau^3 + N^{-4} \right].
$$

Step 6. Discrete Gronwall Lemma. The above two estimates together with (7.45) and (7.46) imply

$$
\begin{aligned}
&\frac{1}{2} D^1 \left\| \nabla\theta_q^{k+1} \right\|_0^2 + \frac{1}{2} D^1 \left\| \nabla\theta_p^{k+1} \right\|_0^2 + D^1 \left| e_r^{k+1} \right|^2 \\
&\quad \leq \tau C \left[\left\| \nabla\theta_p^{k+1} \right\|_0^2 + \left\| \nabla\theta_p^k \right\|_0^2 + \left\| \nabla\theta_q^{k+1} \right\|_0^2 + \left\| \nabla\theta_q^k \right\|_0^2 + \left| e_r^{k+1} \right|^2 + \left| e_r^k \right|^2 \right] + C \left[\tau^4 + N^{-4} \right].
\end{aligned}
$$

Therefore, by the use of Gronwall's Lemma, we can conclude that

$$
\frac{1}{2} \left\| \nabla\theta_q^{k+1} \right\|_0^2 + \frac{1}{2} \left\| \nabla\theta_p^{k+1} \right\|_0^2 + \left| e_r^{k+1} \right|^2 \leq C \exp\left(\left[1 - C\tau \right]^{-1} t^{k+1} \right) \left(\tau^4 + N^{-4} \right).
$$

□

7.6 Numerical experiments

In this section we numerically confirm our theoretical convergence result given in Theorem 56 and illustrate the long time energy conservation of the SAV method. In the following, the

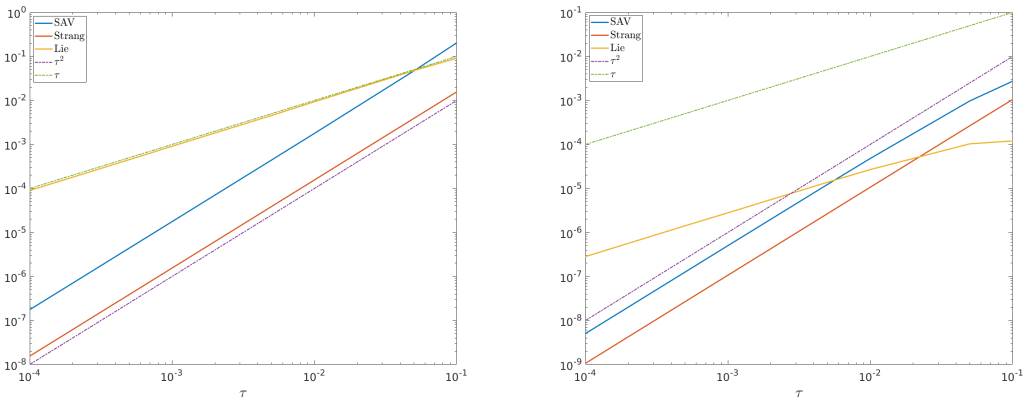


Figure 7.1 – Error e_u (left) and e_H (right) versus step size τ at time $T = 10$.

numerical results have been obtained from Algorithm 2. However, we want to emphasize that Algorithm 1 leads to the same results and has a comparable computational cost for $d = 1$.

Our numerical findings suggest the favorable energy preservation of the SAV method compared to classical splitting methods in certain applications such as for non-linearities with non-integer exponents which arise for instance in context of optical dark and power law solitons with surface plasmonic interactions [65]. For the comparison we use the classical first order Lie and second order Strang splitting which are known for their near energy preservation over long times, see, e.g., [83].

In the numerical examples we plot the deviation of the exact Hamiltonian and the modified Hamiltonian $\tilde{H}$, i.e., $e_{\tilde{H}} = \left|H(t^k) - \tilde{H}^k\right|$, the error between the exact Hamiltonian and the discrete non-modified Hamiltonian $e_H = \left|H(u(t^k)) - H(U^k)\right|$, as well as the L^2 error $e_u = \left\|\left|U^k\right| - \left|u(t^k)\right|\right\|_{L^2(\Omega)}$. We choose the potential $V = 0$ in the Schrödinger equation (7.1).

7.6.1 First test case: cubic nonlinearity

In a first example we consider the nonlinear Schrödinger equation (7.1) with a cubic nonlinearity i.e.

$$f(|u|^2) = \beta\,|u|^2$$

on the spatial domain $\Omega = [-32, 32]$. In Figure 7.1 we choose a mesh size $h = 1/32$ and approximate the soliton solution [26, 18]

$$u(x,t) = \frac{a}{\sqrt{-\beta}}\operatorname{sech}\left(a(x - vt)\right)\exp\left(ivx - 0.5(v^2 - a^2)t\right),$$

with the parameters $a = 1$, $\beta = -1$ and $v = 1$ up to $T = 10$. Figure 7.1 numerically confirms the second-order convergence of the SAV method. The numerical findings also suggest that the error constant of the Strang splitting method is slightly better than the one of the SAV method in this example. In Figure 7.2 we simulate the solitary wave

$$u(t,x) = \frac{\sqrt{2}e^{it}}{\cosh(x)}$$

on the domain $[-\frac{\pi}{0.11}, \frac{\pi}{0.11}]$ with $N = 256$ collocation points and time step size $\tau = 0.01$. We illustrate the evolution of the errors e_H, e_u and $e_{\tilde{H}}$ over long times, i.e., up to $T = 1000$. Our numerical findings confirm the conservation of the modified Hamiltonian by the SAV method, see Figure 7.2. We also observe that the SAV method preserves well the exact energy and L^2 norm over long times. Even though, the error e_H of the Strang splitting seems favorable in this example, we have to stress that the modified Hamiltonian is closer to the value of the real Hamiltonian (see error $e_{\tilde{H}}$ on Figure 7.2).

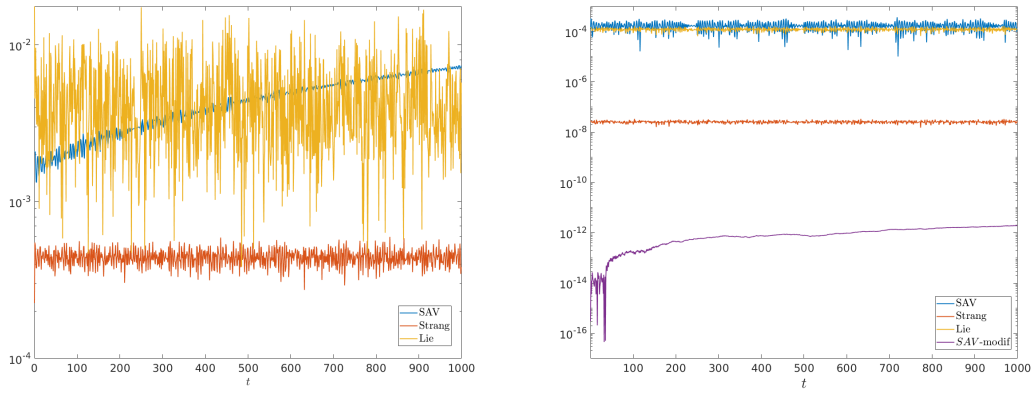


Figure 7.2 – Left Figure: error e_u (left) through time. Right Figure: e_H (blue, red, yellow) and $e_{\tilde{H}}$ (purple) through time.

7.6.2 Second test case: cubic nonlinearity with non-smooth initial condition

In this example we analyse the error behaviour of the SAV scheme in case of non-smooth initial data. For this purpose we solve the NLS equation with cubic nonlinearity with initial data of various regularity. More precisely, we choose $f(|u|^2) = \beta |u|^2$ with $\beta = 1$ and consider $u^0 \in H^\alpha$ with $\alpha = 3/2, 2, 3, 5$ on the spatial domain $\Omega = [-\pi, \pi]$ with $N = 1024$ gridpoints. The discrete initial data of various regularity is generated as proposed in [159].

Figure 7.3 shows the convergence behaviour of the SAV scheme, and the two splitting methods for the initial data of different regularity. We find that if $\alpha < 3$, the SAV method does not maintain its second order convergence rate and for $\alpha = 2$, the SAV scheme reduces to first order. Decreasing the regularity of the initial condition even more, the convergence worsens and becomes less than order 1. A similar order reduction is observed for the splitting schemes, however, for the latter the error starts to oscillate for $\alpha < 3$. Again, the error e_H is favorable for the Strang splitting for all α. However, the modified Hamiltonian is closer to the real Hamiltonian (see Figure 7.4 for $\alpha = \frac{2}{3}$).

7.6.3 Third test case: non-integer exponent

In this example we consider the periodic nonlinear Schrödinger equation (7.1) with nonlinearities with non-integer exponents ([65])

$$f(|u|^2) = \beta |u|^{4/\gamma}, \quad \gamma > 0$$

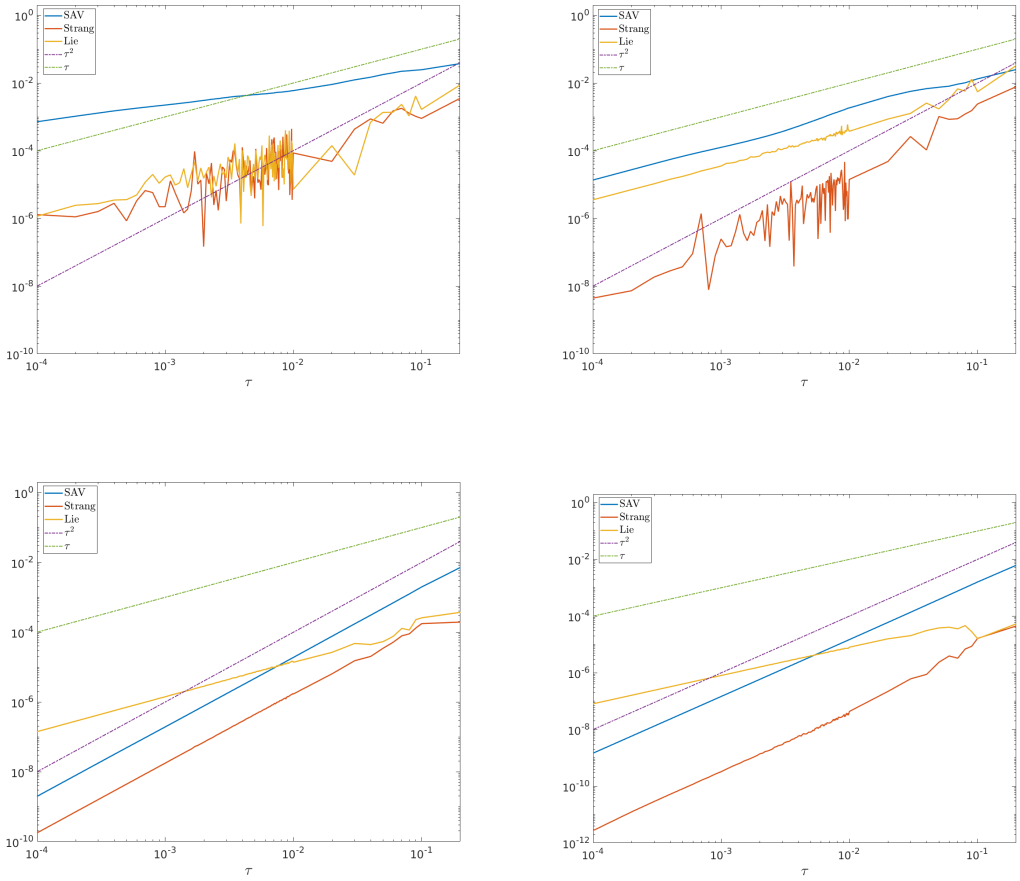


Figure 7.3 – Error e_H versus step size τ for the nonlinear Schrödinger equation starting from different initial conditions in H^α ($\alpha = \frac{2}{3}$ top-left, $\alpha = 2$ top-right, $\alpha = 3$ bottom left and $\alpha = 5$ bottom-right). The dotted lines represent order τ (green) and τ^2 (purple), respectively.

where the Hamiltonian takes the form

$$H(u) = \int_\Omega \frac{1}{2} |\nabla u|^2 + \beta \frac{\gamma}{(4+2\gamma)} |u|^{\frac{4}{\gamma}+2} \, \mathrm{d}x.$$

We carry out simulations for various exponents $\gamma = 2, 8/3, 4, 8$ up to time $T = 10$ with smooth initial value

$$u(0, x) = \sin(x) \in C^\infty([-\pi, \pi]).$$

The error e_H for different exponents γ is plotted in Figure 7.5. Our numerical findings suggest that as γ increases the splitting methods suffer from sever order reduction. This loss of convergence of splitting methods was also observed in [159]. The SAV method, on the other hand, retains its second order energy convergence for non-integer exponents.

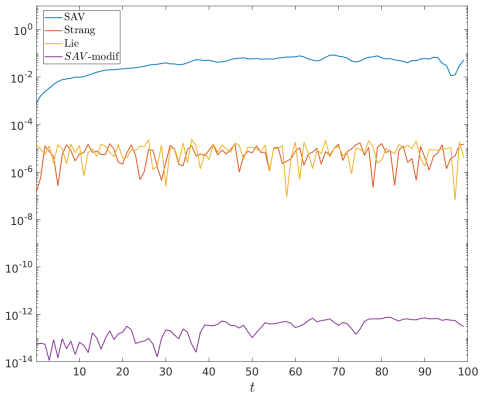


Figure 7.4 – Error e_H (blue, red, yellow) through time and $e_{\tilde{H}}$ (purple) through time for initial condition in $H^{\frac{2}{3}}$.

7.6.4 Computing ground states

We use the SAV scheme to simulate ground states of Bose-Einstein condensates, however, unlike the work of Antoine *et al.* [18], we propose here to use a different strategy. Indeed, in [18], the authors use the SAV scheme presented in Section 7.2 and observe the capacity of the scheme to preserve the initial mass and Hamiltonian for various strengths of the nonlinearity. In the present work, we propose to use a different method and compare the numerical results with reference methods that are designed to simulate the stationary states of the NLS equation for large nonlinearities.

To do so, we reformulate the problem into the solving of a gradient flow equation to compute these stationary states: this method is known has the *gradient flow with discrete normalization* method [20, 23].

The SAV scheme is well adapted to this formulation since its original purpose was the simulation of the gradient flow equations. Details of the reformulation and the adaptation of the SAV scheme to the case can be found in Appendix 7.A.

We here present numerical results obtained choosing $d = 1$, $V(x) = x^2/2$, $\beta = 400$, and $u^0(x) = \frac{\exp(-x^2/2)}{\pi^{1/4}} / \left\|u^0\right\|_0$. We validate our results with the Backward Euler PseudoSpectral (BEPS in short) scheme implemented in the GPELab code [16, 17].

We denote by $\mathcal{E}(u)$ the energy associated to the renormalized system and $\tilde{\mathcal{E}}(u)$ its modified SAV energy (see Appendix 7.A for details).

Figure 7.6 (left) compares the stationary states obtained with the two schemes for $h = 1/8$ in space. We clearly see that they both reach the same steady state. Figure 7.6 (right) depicts the evolution of the energy during the simulation. We observe that the SAV scheme preserves the monotonic decay of the energy. The steady state reached at the end of the simulation has an energy $\tilde{\mathcal{E}}(\phi) \approx 22.90$. However, using the solution ϕ_g obtained for the SAV scheme and computing the "real" energy, we obtain $\mathcal{E}(\phi) \approx 21.36$ which is the value obtained with the BEPS scheme.

We numerically evaluate the order of convergence in space of the SAV scheme for the simulation of ground states. We choose our reference solution to be the result of a simulation with $h = 1/32$. Then, we vary h from $1/2$ to $1/16$. Figure 7.7 shows that the scheme remains second order convergent in space as predicted by our error analysis.

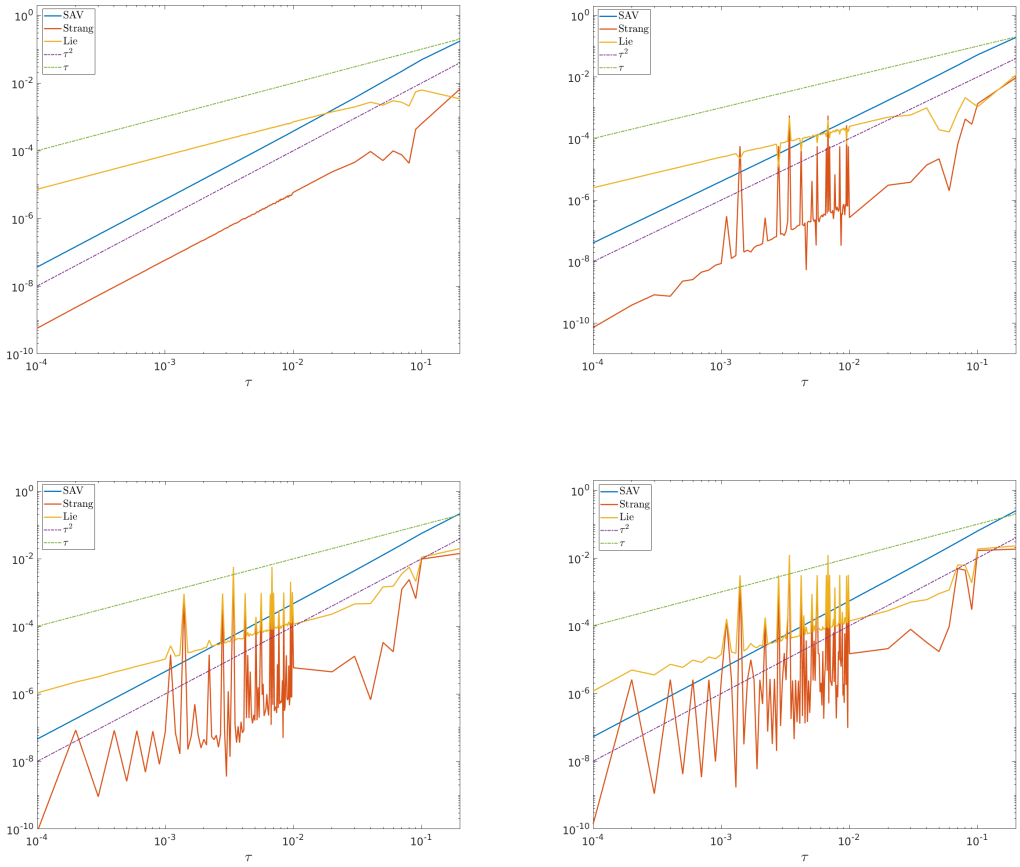


Figure 7.5 – Error e_H versus step size τ for the nonlinear Schrödinger equation with different non-integer exponents ($\gamma = 2$ top-left, $\gamma = \frac{8}{3}$ top-right, $\gamma = 4$ bottom left and $\gamma = 8$ bottom-right). The dotted lines represent order τ (green) and τ^2 (purple), respectively.

7.A Gradient flow with discrete normalization for computing ground state

A common method to compute stationary states of the NLS equation (7.1) with a cubic nonlinearity is to write

$$u(t,x) = \phi(x)\exp(-i\mu t),$$

where μ is defined as the chemical potential of the condensate

$$\mu(\phi) = \int_\Omega \left(\frac{1}{2}|\nabla\phi| + \beta|\phi|^4 + V(x)|\phi|^2\right) \mathrm{d}x.$$

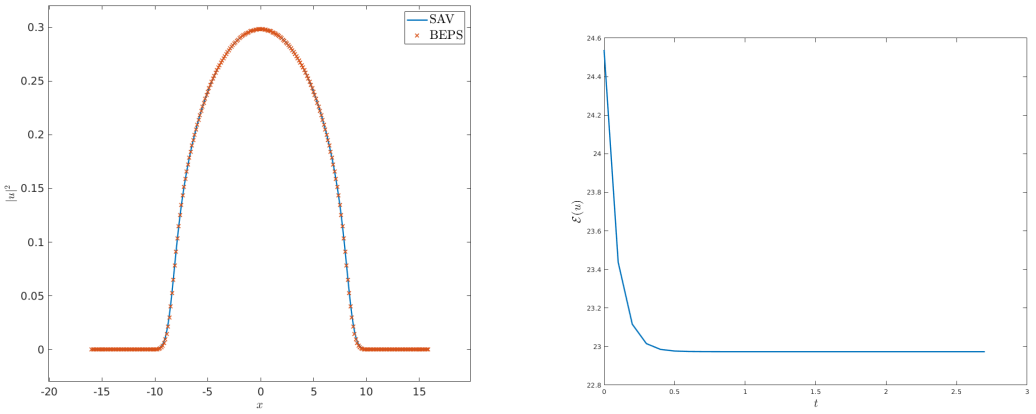


Figure 7.6 – (Left) Comparison of stationary solutions of the NLS equation with a large cubic nonlinearity obtained with the SAV scheme (blue) and the BEPS scheme from GPELab (red). (Right) Evolution of the energy $\tilde{\mathcal{E}}(u)$ for the SAV scheme during the simulation.

Therefore, using the previous reformulation in Equation (7.1), we obtain

$$\mu\phi(x) = -\frac{1}{2}\Delta\phi(x) + \beta\left|\phi(x)\right|^2\phi(x) + V(x)\phi(x).$$

Denoting by $S = \{\phi | \, \|\phi\|_{L^2(\Omega)} = 1\}$ the unit sphere, the ground state $\phi_g \in S$ of the Bose-Einstein condensate is then defined by the solution minimizing the energy functional

$$\mathcal{E}(\phi) = \int_\Omega \left(\frac{1}{2}\left|\nabla\phi\right| + \frac{1}{2}\beta\left|\phi\right|^4 + V(x)\left|\phi\right|^2\right) \mathrm{d}x < +\infty.$$

For the proof of the existence of such state and other mathematical properties we refer the reader to [22].

In the following, we adapt the Scalar Auxiliary Variable method to compute the stationary solutions of Equation (7.1). Therefore, endowing the equation with the normalization constraint, and using the projected gradient method [20], the complete system reads

$$\begin{cases} \partial_t\phi = \frac{1}{2}\Delta\phi - V(x)u - \beta\left|u\right|^2 u(t,x), \\ \|\phi\|^2_{L^2(\Omega)} = 1. \end{cases}$$

Our SAV scheme can be easily adapted to this case, leading to the discrete system

$$\begin{cases} \frac{\phi^+ - \phi^k}{\tau} = \frac{1}{2}D^{(2)}\phi^{k+1/2} - r^{k+1/2}\tilde{G}^{k+1/2}, \\ r^{k+1} - r^k = \frac{1}{2}\left(\tilde{G}^{k+1/2}, \phi^+ - \phi^k\right) \\ \phi^{k+1} = \frac{\phi^+}{\|\phi^+\|^2_L(\Omega)}, \end{cases}$$

with $\phi^{k+1/2} = \frac{\phi^+ + \phi^k}{2}$, $\tilde{G}^{k+1/2}$ a second order approximation of $\frac{\delta\mathcal{E}_1[t^{k+1/2}]}{\delta\phi^{k+1/2}}$. We precise that the

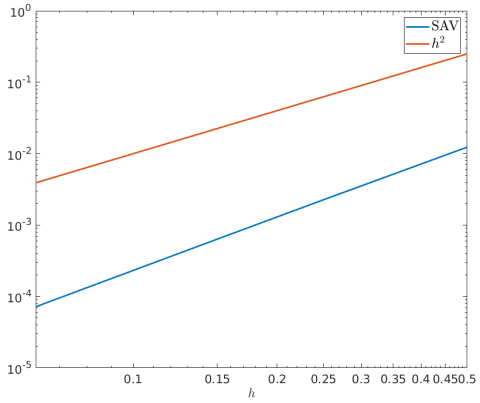


Figure 7.7 – Error e_u versus grid size h for the simulation of ground states with a large cubic nonlinearity.

associated modified SAV energy is

$$\tilde{\mathcal{E}}(\phi) = \int_\Omega \frac{1}{2} |\nabla \phi| \, \mathrm{d}x + r(t)^2 < +\infty.$$

Both Algorithm 1 and Algorithm 2 from Section 7.2 can be applied to compute the solution of the SAV system. Furthermore, using the same calculation as in Section 7.3, we can easily prove that the scheme dissipates the energy and preserves the normalization constraint.

Bibliography

[1] H. Abels and E. Feireisl. "On a diffuse interface model for a two-phase flow of compressible viscous fluids". In: *Indiana Univ. Math. J.* 57.2 (2008), pp. 659–698.

[2] H. Abels and M. Wilke. "Convergence to equilibrium for the Cahn-Hilliard equation with a logarithmic free energy". In: *Nonlinear Anal.* 67.11 (2007), pp. 3176–3193.

[3] R. A. Adams. *Sobolev spaces*. Pure and Applied Mathematics, Vol. 65. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1975, pp. xviii+268.

[4] R. A. Adams and J. J. F. Fournier. *Sobolev spaces*. Second. Vol. 140. Pure and Applied Mathematics (Amsterdam). Elsevier/Academic Press, Amsterdam, 2003, pp. xiv+305.

[5] L. Adenis et al. "Experimental and modeling study of the formation of cell aggregates with differential substrate adhesion". In: *PLOS ONE* 15.2 (Feb. 2020), pp. 1–19.

[6] A. Agosti, S. Marchesi, G. Scita, and P. Ciarletta. "Modelling cancer cell budding in-vitro as a self-organised, non-equilibrium growth process". In: *J. Theoret. Biol.* 492 (2020), pp. 110203, 8.

[7] A. Agosti. "Discontinuous Galerkin finite element discretization of a degenerate Cahn-Hilliard equation with a single-well potential". In: *Calcolo* 56.2 (2019), Paper No. 14, 47.

[8] A. Agosti, P. F. Antonietti, P. Ciarletta, M. Grasselli, and M. Verani. "A Cahn-Hilliard-type equation with application to tumor growth dynamics". In: *Math. Methods Appl. Sci.* 40.18 (2017), pp. 7598–7626.

[9] A. Agosti, C. Cattaneo, C. Giverso, D. Ambrosi, and P. Ciarletta. "A computational framework for the personalized clinical treatment of glioblastoma multiforme". In: *Z. Angew. Math. Mech.* 98.12 (2018), pp. 2307–2327.

[10] A. Agosti, C. Giverso, E. Faggiano, A. Stamm, and P. Ciarletta. "A personalized mathematical tool for neuro-oncology: A clinical case study". In: *Int. J. Non Linear Mech.* 107 (2018), pp. 170–181.

[11] J. Alfonso et al. "The biology and mathematical modelling of glioma invasion: a review". In: *J. R. Soc. Interface* 14.136 (2017), p. 20170490.

[12] L. Almeida, F. Bubba, B. Perthame, and C. Pouchol. "Energy and implicit discretization of the Fokker-Planck and Keller-Segel type equations". In: *Netw. Heterog. Media* 14.1 (2019), pp. 23–41.

[13] L. Almeida, G. Estrada-Rodriguez, L. Oliver, D. Peurichard, A. Poulain, and F. Vallette. "Treatment-induced shrinking of tumour aggregates: A nonlinear volume-filling chemotactic approach". In: *arXiv preprint arXiv:2007.12454* (2020).

[14] W. Alt. "Biased random walk models for chemotaxis and related diffusion approximations". In: *J. Math. Biol.* 9.2 (1980), pp. 147–177.

[15] X. Antoine, W. Bao, and C. Besse. "Computational methods for the dynamics of the nonlinear Schrödinger/Gross-Pitaevskii equations". In: *Comput. Phys. Commun.* 184.12 (2013), pp. 2621–2633.

[16] X. Antoine and R. Duboscq. "GPELab, a Matlab toolbox to solve Gross–Pitaevskii equations I: Computation of stationary solutions". In: *Computer Physics Communications* 185.11 (2014), pp. 2969–2991.

[17] X. Antoine and R. Duboscq. "GPELab, a Matlab toolbox to solve Gross–Pitaevskii equations II: Dynamics and stochastic simulations". In: *Computer Physics Communications* 193 (2015), pp. 95–117.

[18] X. Antoine, J. Shen, and Q. Tang. "Scalar Auxiliary Variable/Lagrange multiplier based pseudospectral schemes for the dynamics of nonlinear Schrödinger/Gross-Pitaevskii equations". In: *J. Comput. Phys.* 437 (2021), p. 110328.

[19] K. Baba and M. Tabata. "On a conservative upwind finite element scheme for convective diffusion equations". In: *RAIRO Anal. Numér.* 15.1 (1981), pp. 3–25.

[20] W. Bao. "Ground states and dynamics of multicomponent Bose-Einstein condensates". In: *Multiscale Model. Simul.* 2.2 (2004), pp. 210–236.

[21] W. Bao. "The nonlinear Schrödinger equation and applications in Bose-Einstein condensation and plasma physics". In: *Dynamics in models of coarsening, coagulation, condensation and quantization.* Vol. 9. Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap. World Sci. Publ., Hackensack, NJ, 2007, pp. 141–239.

[22] W. Bao and Y. Cai. "Mathematical theory and numerical methods for Bose-Einstein condensation". In: *Kinet. Relat. Models* 6.1 (2013), pp. 1–135.

[23] W. Bao and Q. Du. "Computing the ground state solution of Bose-Einstein condensates by a normalized gradient flow". In: *SIAM J. Sci. Comput.* 25.5 (2004), pp. 1674–1697.

[24] W. Bao, D. Jaksch, and P. A. Markowich. "Numerical solution of the Gross-Pitaevskii equation for Bose-Einstein condensation". In: *J. Comput. Phys.* 187.1 (2003), pp. 318–342.

[25] W. Bao, S. Jin, and P. A. Markowich. "On time-splitting spectral approximations for the Schrödinger equation in the semiclassical regime". In: *J. Comput. Phys.* 175.2 (2002), pp. 487–524.

[26] W. Bao, Q. Tang, and Z. Xu. "Numerical methods and comparison for computing dark and bright solitons in the nonlinear Schrödinger equation". In: *J. Comput. Phys.* 235 (2013), pp. 423–445.

[27] L. Barazzuol et al. "In vitro evaluation of combined temozolomide and radiotherapy using X rays and high-linear energy transfer radiation for glioblastoma". In: *Radiat. Res.* 177.5 (2012), pp. 651–662.

[28] J. W. Barrett, J. F. Blowey, and H. Garcke. "Finite element approximation of the Cahn-Hilliard equation with degenerate mobility". In: *SIAM J. Numer. Anal.* 37.1 (1999), pp. 286–318.

[29] M. Ben Amar and A. Goriely. "Growth and instability in elastic tissues". In: *J. Mech. Phys. Solids* 53.10 (2005), pp. 2284–2319.

[30] E. Ben-Jacob, I. Cohen, and H. Levine. "Cooperative self-organization of microorganisms". In: *Adv. Phys.* 49.4 (2000), pp. 395–554.

[31] C. Besse, B. Bidégaray, and S. Descombes. "Order estimates in time of splitting methods for the nonlinear Schrödinger equation". In: *SIAM J. Numer. Anal.* 40.1 (2002), pp. 26–40.

[32] M. Bessemoulin-Chatard and A. Jüngel. "A finite volume scheme for a Keller-Segel model with additional cross-diffusion". In: *IMA J. Numer. Anal.* 34.1 (2014), pp. 96–122.

[33] A. Blanchet, J. Dolbeault, and B. Perthame. "Two-dimensional Keller-Segel model: optimal critical mass and qualitative properties of the solutions". In: *Electron. J. Differential Equations* (2006), No. 44, 32.

[34] A. Blanchet and P. Laurençot. "The parabolic-parabolic Keller-Segel system with critical diffusion as a gradient flow in $\mathbb{R}^d$, $d \geq 3$". In: *Comm. Partial Differential Equations* 38.4 (2013), pp. 658–686.

[35] J. F. Blowey and C. M. Elliott. "The Cahn-Hilliard gradient theory for phase separation with nonsmooth free energy. I. Mathematical analysis". In: *European J. Appl. Math.* 2.3 (1991), pp. 233–280.

[36] A. Bouchriti, M. Pierre, and N. E. Alaa. "Remarks on the asymptotic behavior of scalar auxiliary variable (SAV) schemes for gradient-like flows". In: *J. Appl. Anal. Comput.* 10.5 (2020), pp. 2198–2219.

[37] F. Boyer and C. Lapuerta. "Study of a three component Cahn-Hilliard flow model". In: *M2AN Math. Model. Numer. Anal.* 40.4 (2006), pp. 653–687.

[38] F. Boyer and S. Minjeaud. "Hierarchy of consistent n-component Cahn-Hilliard systems". In: *Math. Models Methods Appl. Sci.* 24.14 (2014), pp. 2885–2928.

[39] F. Boyer and S. Minjeaud. "Numerical schemes for a three component Cahn-Hilliard model". In: *ESAIM Math. Model. Numer. Anal.* 45.4 (2011), pp. 697–738.

[40] M. Brandao, T. Simon, G. Critchley, and G. Giamas. "Astrocytes, the rising stars of the glioblastoma microenvironment". In: *Glia* 67.5 (2019), pp. 779–790.

[41] A. Bray. "Theory of phase-ordering kinetics". In: *Adv. Phys.* 43.3 (1994), pp. 357–459.

[42] S. C. Brenner, S. Gu, T. Gudi, and L.-y. Sung. "A quadratic C° interior penalty method for linear fourth order boundary value problems with boundary conditions of the Cahn-Hilliard type". In: *SIAM J. Numer. Anal.* 50.4 (2012), pp. 2088–2110.

[43] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*. Third. Vol. 15. Texts in Applied Mathematics. Springer, New York, 2008, pp. xviii+397.

[44] F. Bubba, T. Lorenzi, and F. R. Macfarlane. "From a discrete model of chemotaxis with volume-filling to a generalized Patlak-Keller-Segel model". In: *Proc. A.* 476.2237 (2020), pp. 20190871, 19.

[45] F. Bubba et al. "A chemotaxis-based explanation of spheroid formation in 3D cultures of breast cancer cells". In: *J. Theor. Biol.* 479 (2019), pp. 73–80.

[46] H. Byrne and L. Preziosi. "Modelling solid tumour growth using the theory of mixtures". In: *Math. Med. Biol.* 20.4 (2003), pp. 341–366.

[47] J. W. Cahn. "On spinodal decomposition". In: *Acta metall.* 9.9 (Sept. 1961), pp. 795–801.

[48] J. W. Cahn and J. E. Hilliard. "Free Energy of a Nonuniform System. I. Interfacial Free Energy". en. In: *J. Chem. Phys.* 28.2 (Feb. 1958), pp. 258–267.

[49] W. Cai, C. Jiang, Y. Wang, and Y. Song. "Structure-preserving algorithms for the two-dimensional sine-Gordon equation with Neumann boundary conditions". In: *J. Comput. Phys.* 395 (2019), pp. 166–185.

[50] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral methods*. Scientific Computation. Fundamentals in single domains. Springer-Verlag, Berlin, 2006, pp. xxii+563.

[51] X. Cao. "Global bounded solutions of the higher-dimensional Keller-Segel system under smallness conditions in optimal spaces". In: *Discrete Contin. Dyn. Syst.* 35.5 (2015), pp. 1891–1904.

[52] J. A. Carrillo, S. Hittmeir, and A. Jüngel. "Cross diffusion and nonlinear diffusion preventing blow up in the Keller-Segel model". In: *Math. Models Meth. Appl. Sci.* 22.12 (2012), p. 1250041.

[53] V. Chalupecky. "Numerical studies of Cahn-Hilliard equation and applications in image processing". In: *Proceedings of Czech–Japanese Seminar in Applied Mathematics*. 2004.

[54] M. A. J. Chaplain and G. Lolas. "Mathematical modelling of cancer invasion of tissue: dynamic heterogeneity". In: *Netw. Heterog. Media* 1.3 (2006), pp. 399–439.

[55] C. Chatelain, T. Balois, P. Ciarletta, and M. Ben Amar. "Emergence of microstructural patterns inskin cancer: a phase separation analysis in a binary mixture". In: *New J. Phys.* 13 (2011), pp. 339–357.

[56] C. Chatelain, P. Ciarletta, and M. Ben Amar. "Morphological changes in early melanoma development: influence of nutrients, growth inhibitors and cell-adhesion mechanisms". In: *J. Theoret. Biol.* 290 (2011), pp. 46–59.

[57] X. Chen et al. "A feedforward mechanism mediated by mechanosensitive ion channel PIEZO1 and tissue mechanics promotes glioma aggression". In: *Neuron* 100.4 (2018), pp. 799–815.

[58] L. Cherfils, E. Feireisl, M. Michálek, A. Miranville, M. Petcu, and D. Pražák. "The compressible Navier-Stokes-Cahn-Hilliard equations with dynamic boundary conditions". In: *Math. Models Methods Appl. Sci.* 29.14 (2019), pp. 2557–2584.

[59] L. Cherfils, A. Miranville, and S. Zelik. "On a generalized Cahn-Hilliard equation with biological applications". In: *Discrete Contin. Dyn. Syst. Ser. B* 19.7 (2014), pp. 2013–2026.

[60] L. Cherfils, A. Miranville, and S. Zelik. "The Cahn-Hilliard equation with logarithmic potentials". In: *Milan J. Math.* 79.2 (2011), pp. 561–596.

[61] P. Ciarletta, L. Foret, and M. Ben Amar. "The radial growth phase of malignant melanoma : muti-phase modelling, numerical simulation and linear stability". In: *J. R. Soc. Interface* 8.56 (2011), pp. 345–368.

[62] D. S. Cohen and J. D. Murray. "A generalized diffusion model for growth and dispersal in a population". In: *J. Math. Biol.* 12.2 (1981), pp. 237–249.

[63] M. C. Colombo, C. Giverso, E. Faggiano, C. Boffano, F. Acerbi, and P. Ciarletta. "Towards the Personalized Treatment of Glioblastoma: Integrating Patient-Specific Clinical Data in a Continuous Mechanical Model". In: *PLoS ONE* 10 (2015).

[64] J. Condeelis, R. H. Singer, and J. E. Segall. "The great escape: when cancer cells hijack the genes for chemotaxis and motility". In: *Annu. Rev. Cell Dev. Biol.* 21 (2005), pp. 695–718.

[65] S. H. Crutcher and A. Osei. "Derivation of the Effective Nonlinear Schrödinger Equations for Dark and Power Law Spatial Plasmon-Polariton Solitons Using Nano Self-Focusing". In: *Progress In Electromagnetics Research* 29 (2011). Publisher: EMW Publishing, pp. 83–103.

[66] H. Darcy. *Les fontaines publiques de la ville de Dijon: exposition et application...* Victor Dalmont, 1856.

[67] E. Davoli, H. Ranetbauer, L. Scarpa, and L. Trussardi. "Degenerate nonlocal Cahn-Hilliard equations: well-posedness, regularity and local asymptotics". In: *Ann. Inst. H. Poincaré Anal. Non Linéaire* 37.3 (2020), pp. 627–651.

[68] S. L. Di Jia, D. Li, H. Xue, D. Yang, and Y. Liu. "Mining TCGA database for genes of prognostic value in glioblastoma microenvironment". In: *Aging (Albany NY)* 10.4 (2018), p. 592.

[69] Y. Dolak and C. Schmeiser. "The Keller–Segel model with logistic sensitivity function and small diffusivity". In: *SIAM J. Appl. Math.* 66.1 (2005), pp. 286–308.

[70] D. Dormann and C. J. Weijer. "Chemotactic cell movement during Dictyostelium development and gastrulation". In: *Curr. Opin. Genet. Dev.* 16.4 (2006), pp. 367–373.

[71] M. Ebenbeck and H. Garcke. "Analysis of a Cahn-Hilliard-Brinkman model for tumour growth with chemotaxis". In: *J. Differential Equations* 266.9 (2019), pp. 5998–6036.

[72] M. Ebenbeck, H. Garcke, and R. Nürnberg. "Cahn-Hilliard-Brinkman systems for tumour growth". In: *arXiv preprint arXiv:2003.08314* (2020).

[73] C. Eck, H. Garcke, and P. Knabner. *Mathematical modeling*. Springer Undergraduate Mathematics Series. Springer, Cham, 2017, pp. xv+509.

[74] C. M. Elliott. "The Cahn-Hilliard model for the kinetics of phase separation". In: *Mathematical models for phase change problems (Óbidos, 1988)*. Vol. 88. Internat. Ser. Numer. Math. Birkhäuser, Basel, 1989, pp. 35–73.

[75] C. M. Elliott, D. A. French, and F. A. Milner. "A second order splitting method for the Cahn-Hilliard equation". In: *Numer. Math.* 54 (1989), pp. 575–590.

[76] C. M. Elliott and H. Garcke. "On the Cahn-Hilliard Equation with Degenerate Mobility". In: *SIAM J. Math. Anal.* 27.2 (Mar. 1996), pp. 404–423.

[77] C. M. Elliott and Z. Songmu. "On the Cahn-Hilliard equation". In: *Arch. Rational Mech. Anal.* 96.4 (1986), pp. 339–357.

[78] Y. Epshteyn and A. Izmirlioglu. "Fully discrete analysis of a discontinuous finite element method for the Keller-Segel chemotaxis model". In: *J. Sci. Comput.* 40.1-3 (2009), pp. 211–256.

[79] Y. Epshteyn and A. Kurganov. "New interior penalty discontinuous Galerkin methods for the Keller-Segel chemotaxis model". In: *SIAM J. Numer. Anal.* 47.1 (2008/09), pp. 386–408.

[80] J. L. Ericksen. "Liquid crystals with variable degree of orientation". In: *Arch. Rational Mech. Anal.* 113.2 (1990), pp. 97–120.

[81] J. Erlebacher, M. J. Aziz, A. Karma, N. Dimitrov, and K. Sieradzki. "Evolution of nanoporosity in dealloying". In: *Nature* 410.6827 (Mar. 2001). Number: 6827 Publisher: Nature Publishing Group, pp. 450–453.

[82] D. J. Eyre. "An Unconditionally Stable One-Step Scheme for Gradient Systems". 1997.

[83] E. Faou. *Geometric numerical integration and Schrödinger equations*. Zurich Lectures in Advanced Mathematics. European Mathematical Society (EMS), Zürich, 2012, pp. xiii+138.

[84] X. Feng, B. Li, and S. Ma. "High-order Mass- and Energy-conserving SAV-Gauss Collocation Finite Element Methods for the Nonlinear Schrödinger Equation". In: *SIAM Journal on Numerical Analysis* 59.3 (2021), pp. 1566–1591.

[85] M. E. Fernandez-Sanchez et al. "Mechanical induction of the tumorigenic β-catenin pathway by tumour growth pressure". In: *Nature* 523.7558 (2015), pp. 92–95.

[86] F. Filbet. "A finite volume scheme for the Patlak-Keller-Segel chemotaxis model". In: *Numer. Math.* 104.4 (2006), pp. 457–488.

[87] B. Fornberg. *A practical guide to pseudospectral methods*. Vol. 1. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, 1996, pp. x+231.

[88] Y. A. Fouad and C. Aanei. "Revisiting the hallmarks of cancer". In: *Am. J. Cancer Res.* 7.5 (2017), p. 1016.

[89] H. B. Frieboes et al. "Computer simulation of glioma growth and morphology". In: *Neuroimage* 37 (2007), S59–S70.

[90] H. B. Frieboes, F. Jin, Y.-L. Chuang, S. M. Wise, J. S. Lowengrub, and V. Cristini. "Three-dimensional multispecies nonlinear tumor growth—II: Tumor invasion and angiogenesis". In: *J. Theor. Biol.* 264.4 (2010), pp. 1254–1278.

[91] P. Friedl. "Prespecification and plasticity: shifting mechanisms of cell migration". In: *Curr. Opin. Cell Biol.* 16.1 (2004), pp. 14–23.

[92] S. Frigeri, K. F. Lam, E. Rocca, and G. Schimperna. "On a multi-species Cahn-Hilliard-Darcy tumor growth model with singular potentials". In: *Commun. Math. Sci.* 16.3 (2018), pp. 821–856.

[93] Y. Fu, W. Cai, and Y. Wang. "A structure-preserving algorithm for the fractional nonlinear Schrödinger equation based on the SAV approach". In: *arXiv preprint arXiv:1911.07379* (2019).

[94] H. Fujii. "Some remarks on finite element analysis of time dependent field problems". In: *Theorie and Practice in Finite Element Structural Analysis (Y. Yamada and R.H. Gallager editions)*. Univ. Tokyo Press, 1973, pp. 91–106.

[95] C. G. Gal and M. Grasselli. "Asymptotic behavior of a Cahn-Hilliard-Navier-Stokes system in 2D". In: *Ann. Inst. H. Poincaré Anal. Non Linéaire* 27.1 (2010), pp. 401–436.

[96] J. Galon and D. Bruni. "Approaches to treat immune hot, altered and cold tumours with combination immunotherapies". In: *Nature Reviews Drug Discovery* 18.3 (Mar. 2019). Number: 3 Publisher: Nature Publishing Group, pp. 197–218.

[97] H. Garcke, K. F. Lam, R. Nürnberg, and E. Sitka. "A multiphase Cahn-Hilliard-Darcy model for tumour growth with necrosis". In: *Math. Models Methods Appl. Sci.* 28.3 (2018), pp. 525–577.

[98] H. Garcke, K. F. Lam, E. Sitka, and V. Styles. "A Cahn-Hilliard-Darcy model for tumour growth with chemotaxis and active transport". In: *Math. Models Methods Appl. Sci.* 26.6 (2016), pp. 1095–1148.

[99] G. Giacomin and J. L. Lebowitz. "Phase segregation dynamics in particle systems with long range interactions. I. Macroscopic limits". In: *J. Statist. Phys.* 87.1-2 (1997), pp. 37–61.

[100] G. Giacomin and J. L. Lebowitz. “Phase segregation dynamics in particle systems with long range interactions. II. Interface motion”. In: *SIAM J. Appl. Math.* 58.6 (1998), pp. 1707–1729.

[101] G. Gilardi, A. Miranville, and G. Schimperna. “Long time behavior of the Cahn-Hilliard equation with irregular potentials and dynamic boundary conditions”. In: *Chin. Ann. Math. Ser. B* 31.5 (2010), pp. 679–712.

[102] A. Giorgini, M. Grasselli, and H. Wu. “The Cahn-Hilliard-Hele-Shaw system with singular potential”. In: *Ann. Inst. H. Poincaré Anal. Non Linéaire* 35.4 (2018), pp. 1079–1118.

[103] G. H. Golub and C. F. Van Loan. *Matrix computations*. Fourth. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, 2013, pp. xiv+756.

[104] R. A. Gonzales, J. Eisert, I. Koltracht, M. Neumann, and G. Rawitscher. “Integral equation method for the continuous spectrum radial Schrödinger equation”. In: *J. Comput. Phys.* 134.1 (1997), pp. 134–149.

[105] R. A. Gonzales, S.-Y. Kang, I. Koltracht, and G. Rawitscher. “Integral equation method for coupled Schrödinger equations”. In: *J. Comput. Phys.* 153.1 (1999), pp. 160–202.

[106] D. Gottlieb, M. Y. Hussaini, and S. A. Orszag. “Theory and applications of spectral methods”. In: *Spectral methods for partial differential equations (Hampton, Va., 1982)*. SIAM, Philadelphia, PA, 1984, pp. 1–54.

[107] E. P. Gross. “Structure of a quantized vortex in boson systems”. In: *Il Nuovo Cimento (1955-1965)* 20.3 (1961), pp. 454–477.

[108] G. Grün and M. Rumpf. “Nonnegativity preserving convergent schemes for the thin film equation”. In: *Numer. Math.* 87.1 (2000), pp. 113–152.

[109] M. E. Gurtin. “On a nonequilibrium thermodynamics of capillarity and phase”. In: *Quart. Appl. Math.* 47.1 (1989), pp. 129–145.

[110] M. E. Gurtin, E. Fried, and L. Anand. *The mechanics and thermodynamics of continua*. Cambridge University Press, Cambridge, 2010, pp. xxii+694.

[111] R. Gutmann et al. “Interstitial hypertension in head and neck tumors in patients: correlation with tumor size”. In: *Cancer research* 52.7 (1992), pp. 1993–1995.

[112] Y. Han, Z. Li, J. Tao, and M. Ma. “Pattern formation for a volume-filling chemotaxis model with logistic growth”. In: *J. Math. Anal. Appl.* 448.2 (2017), pp. 885–907.

[113] D. Hanahan and R. A. Weinberg. “Hallmarks of cancer: the next generation”. In: *Cell* 144.5 (2011), pp. 646–674.

[114] M. Heida, J. Málek, and K. R. Rajagopal. “On the development and generalizations of Cahn-Hilliard equations within a thermodynamic framework”. In: *Z. Angew. Math. Phys.* 63.1 (2012), pp. 145–169.

[115] C.-H. Heldin, K. Rubin, K. Pietras, and A. Östman. “High interstitial fluid pressure—an obstacle in cancer therapy”. In: *Nature Reviews Cancer* 4.10 (2004), pp. 806–813.

[116] T. Hillen and K. Painter. “Global existence for a parabolic chemotaxis model with prevention of overcrowding”. In: *Adv. in Appl. Math.* 26.4 (2001), pp. 280–301.

[117] T. Hillen and K. J. Painter. “A user’s guide to PDE models for chemotaxis”. In: *J. Math. Biol.* 58.1-2 (2009), pp. 183–217.

[118] D. Horstmann. “The Keller-Segel model in chemotaxis and its consequences”. In: *I, Jahresber. Deutsch. Math.-Verein* (2004), pp. 103–165.

[119] A. Hurwitz. "Ueber die Bedingungen, unter welchen eine Gleichung nur Wurzeln mit negativen reellen Theilen besitzt". In: *Math. Ann.* 46.2 (1895), pp. 273–284.

[120] M. Ibrahim and M. Saad. "On the efficacy of a control volume finite element method for the capture of patterns for a volume-filling chemotaxis model". In: *Comput. Math. with Appl.* 68.9 (2014), pp. 1032–1051.

[121] A. Iuorio and S. Melchionna. "Long-time behavior of a nonlocal Cahn-Hilliard equation with reaction". In: *Discrete Contin. Dyn. Syst.* 38.8 (2018), pp. 3765–3788.

[122] E. F. Keller and L. A. Segel. "Model for chemotaxis". In: *J. Theor. Biol.* 30.2 (1971), pp. 225–234.

[123] E. F. Keller and L. A. Segel. "Initiation of slime mold aggregation viewed as an instability". In: *J. Theor. Biol.* 26.3 (1970), pp. 399–415.

[124] E. F. Keller and L. A. Segel. "Traveling bands of chemotactic bacteria: A theoretical analysis". In: *J. Theor. Biol.* 30.2 (1971), pp. 235–248.

[125] J. S. Kennedy and D. Marsh. "Pheromone-regulated anemotaxis in flying moths". In: *Science* 184.4140 (1974), pp. 999–1001.

[126] E. Khain and L. M. Sander. "A generalized Cahn-Hilliard equation for biological applications". In: *Physical Review E* 77.5 (2008). arXiv: 0801.2574.

[127] J. Kim. "Phase field computations for ternary fluid flows". In: *Comput. Methods Appl. Mech. Eng.* 196.45 (Sept. 2007), pp. 4779–4788.

[128] Y. Kim and S. Kumar. "CD44-mediated adhesion to hyaluronic acid contributes to mechanosensing and invasive motility". In: *Mol. Cancer Res.* 12.10 (2014), pp. 1416–1429.

[129] D. Kuzmin and R. Löhner, eds. *Flux-corrected transport*. Scientific Computation. Principles, algorithms, and applications. Springer-Verlag, Berlin, 2005, pp. xiv+301.

[130] S. Y. Lee. "Temozolomide resistance in glioblastoma multiforme". In: *Genes Dis.* 3.3 (2016), pp. 198–210.

[131] J. Leray. "Sur le mouvement d'un liquide visqueux emplissant l'espace". In: *Acta Math.* 63.1 (1934), pp. 193–248.

[132] J. F. Li and J. Lowengrub. "The effects of cell compressibility, motility and contact inhibition on the growth of tumor cell clusters using the Cellular Potts Model". In: *J. Theor. Biol.* 343 (2014), pp. 79–91.

[133] X. H. Li, C.-W. Shu, and Y. Yang. "Local discontinuous Galerkin method for the Keller-Segel chemotaxis model". In: *J. Sci. Comput.* 73.2-3 (2017), pp. 943–967.

[134] J.-L. Lions. *Quelques méthodes de résolution des problèmes aux limites non linéaires*. Dunod; Gauthier-Villars, Paris, 1969, pp. xx+554.

[135] J.-L. Lions. *Quelques méthodes de résolution des problèmes aux limites non linéaires*. Dunod; Gauthier-Villars, Paris, 1969, pp. xx+554.

[136] P.-L. Lions. *Mathematical topics in fluid mechanics. Vol. 2*. Vol. 10. Oxford Lecture Series in Mathematics and its Applications. Compressible models, Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, 1998, pp. xiv+348.

[137] I. S. Liu. "Method of Lagrange multipliers for exploitation of the entropy principle". In: *Arch. Rational Mech. Anal.* 46 (1972), pp. 131–148.

[138] Q.-X. Liu et al. "Phase separation explains a new class of self-organized spatial patterns in ecological systems". In: *Proc. Natl. Acad. Sci. USA* 110.29 (2013), pp. 11905–11910.

[139] T. Lorenzi, A. Lorz, and B. Perthame. "On interfaces between cell populations with different mobilities". In: *Kinet. Relat. Models* 10.1 (2017), pp. 299–311.

[140] D. N. Louis et al. "The 2016 World Health Organization classification of tumors of the central nervous system: a summary". In: *Acta Neuropathol.* 131.6 (2016), pp. 803–820.

[141] J. Lowengrub and L. Truskinovsky. "Quasi-incompressible Cahn-Hilliard fluids and topological transitions". In: *R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci.* 454.1978 (1998), pp. 2617–2654.

[142] J. Lowengrub, E. Titi, and K. Zhao. "Analysis of a mixture model of tumor growth". In: *European J. Appl. Math.* 24.5 (2013), pp. 691–734.

[143] M. Ma, C. Ou, and Z.-A. Wang. "Stationary solutions of a volume-filling chemotaxis model with logistic growth and their stability". In: *SIAM J. Appl. Math.* 72.3 (2012), pp. 740–766.

[144] G. D. Maurer, D. P. Brucker, and J. P. Steinbach. "Loss of cell-matrix contact increases hypoxia-inducible factor-dependent transcriptional activity in glioma cells". In: *Biochem. Biophys. Res. Commun.* 515.1 (2019), pp. 77–84.

[145] H. Meinhardt. "Models for positional signalling with application to the dorsoventral patterning of insects and segregation into different cell types". In: *Development* 107.Supplement (1989), pp. 169–180.

[146] A. Miranville. "The Cahn-Hilliard equation and some of its variants". In: *AIMS Mathematics* 2 (2017), pp. 479–544.

[147] A. Miranville. *The Cahn-Hilliard equation.* Vol. 95. CBMS-NSF Regional Conference Series in Applied Mathematics. Recent advances and applications. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2019, pp. xiv+216.

[148] I. Müller. *Thermodynamics.* Pitman, 1985.

[149] J. Murray. *Mathematical biology II: spatial models and biomedical applications.* Springer New York, 2001.

[150] B. Nagorcka and J. Mooney. "From stripes to spots: prepatterns which can be produced in the skin by a reaction-diffusion system". In: *Math. Med. Biol* 9.4 (1992), pp. 249–267.

[151] G. A. Narsilio, O. Buzzi, S. Fityus, T. S. Yun, and D. W. Smith. "Upscaling of Navier–Stokes equations in porous media: Theoretical, numerical and experimental approach". In: *Computers and Geotechnics* 36.7 (2009), pp. 1200–1206.

[152] P. A. Netti, L. T. Baxter, Y. Boucher, R. Skalak, and R. K. Jain. "Time-dependent behavior of interstitial fluid pressure in solid tumors: implications for drug delivery". In: *Cancer research* 55.22 (1995), pp. 5451–5458.

[153] H. T. Nia et al. "Solid stress and elastic energy as measures of tumour mechanopathology". In: *Nature biomedical engineering* 1.1 (2016), pp. 1–11.

[154] A. Novick-Cohen. "Chapter 4 The Cahn–Hilliard Equation". en. In: *Handbook of Differential Equations: Evolutionary Equations.* Vol. 4. Elsevier, 2008, pp. 201–228.

[155] K. Oizel et al. "Efficient mitochondrial glutamine targeting prevails over glioblastoma metabolic plasticity". In: *Clin. Cancer Res.* 23.20 (2017), pp. 6292–6304.

[156] L. Oliver, A. Álvarez-Arenas, C. Salaud, et al. "A simple 3D cell culture method for studying the interactions between primary Glioblastoma cells and tumor-activated stromal cells". In: *Under review* (2020).

[157] L. Oliver, L. Lalier, C. Salaud, D. Heymann, P. F. Carton, and F. Vallette. "Drug resistance in Glioblastoma: Are persisters the key to therapy?" In: *Under review* (2020).

[158] K. Osaki and A. Yagi. "Finite dimensional attractor for one-dimensional Keller-Segel equations". In: *Funkcial. Ekvac.* 44.3 (2001), pp. 441–469.

[159] A. Ostermann and K. Schratz. "Low regularity exponential-type integrators for semilinear Schrödinger equations". In: *Found. Comput. Math.* 18.3 (2018), pp. 731–755.

[160] H. G. Othmer and T. Hillen. "The diffusion limit of transport equations II: Chemotaxis equations". In: *SIAM J. Appl. Math.* 62.4 (2002), pp. 1222–1250.

[161] Q. Ouyang and H. L. Swinney. "Transition from a uniform state to hexagonal and striped Turing patterns". In: *Nature* 352.6336 (1991), pp. 610–612.

[162] K. J. Painter and T. Hillen. "Volume-filling and quorum-sensing in models for chemosensitive movement". In: *Can. Appl. Math. Q.* 10.4 (2002), pp. 501–543.

[163] K. Painter. "Modelling cell migration strategies in the extracellular matrix". In: *J. Math. Biol.* 58.4-5 (2009), p. 511.

[164] C. S. Patlak. "Random walk with persistence and external bias". In: *Bull. Math. Biophys.* 15 (1953), pp. 311–338.

[165] B. Perthame and A. Poulain. "Relaxation of the Cahn–Hilliard equation with singular single-well potential and degenerate mobility". In: *European J. Appl. Math.* 32.1 (2021), pp. 89–112.

[166] A. Poulain and F. Bubba. "A nonnegativity preserving scheme for the relaxed Cahn-Hilliard equation with single-well potential and degenerate mobility". In: *ArXiv* (2020).

[167] A. Poulain and F. Bubba. *test A nonnegativity preserving scheme for the relaxed Cahn-Hilliard equation with single-well potential and degenerate mobility*. 2020. arXiv: 1910.13211 [math.AP].

[168] A. Quarteroni and A. Valli. *Numerical approximation of partial differential equations*. Vol. 23. Springer Series in Computational Mathematics. Springer-Verlag, Berlin, 1994, pp. xvi+543.

[169] J. van Rijn, J. J. Heimans, J. van den Berg, P. van der Valk, and B. J. Slotman. "Survival of human glioma cells treated with various combination of temozolomide and X-rays". In: *Int. J. Radiat. Oncol. Biol. Phys.* 47.3 (2000), pp. 779–784.

[170] W. Roos et al. "Apoptosis in malignant glioma cells triggered by the temozolomide-induced DNA lesion O 6-methylguanine". In: *Oncogene* 26.2 (2007), pp. 186–197.

[171] P. G. Saffman. "Viscous fingering in Hele-Shaw cells". In: *Journal of Fluid Mechanics* 173 (1986), pp. 73–94.

[172] N. Saito. "Conservative numerical schemes for the Keller-Segel system and numerical results". In: *Mathematical analysis on the self-organization and self-similarity*. RIMS Kôkyûroku Bessatsu, B15. Res. Inst. Math. Sci. (RIMS), Kyoto, 2009, pp. 125–146.

[173] N. Saito. "Conservative upwind finite-element method for a simplified Keller-Segel system modelling chemotaxis". In: *IMA J. Numer. Anal.* 27.2 (2007), pp. 332–365.

[174] N. Saito. "Error analysis of a conservative finite-element approximation for the Keller-Segel system of chemotaxis". In: *Commun. Pure Appl. Anal.* 11.1 (2012), pp. 339–364.

[175] N. Saito and T. Suzuki. "Notes on finite difference schemes to a parabolic-elliptic system modelling chemotaxis". In: *Appl. Math. Comput.* 171.1 (2005), pp. 72–90.

[176] J. Saragosti, V. Calvez, N. Bournaveas, A. Buguin, P. Silberzan, and B. Perthame. "Mathematical description of bacterial traveling pulses". In: *PLoS Comput. Biol.* 6.8 (2010), e1000890.

[177] W. Sarfaraz and A. Madzvamuse. "Domain-Dependent Stability Analysis of a Reaction–Diffusion Model on Compact Circular Geometries". In: *Int. J. Bifurc. Chaos Appl. Sci. Eng.* 28.08 (2018), p. 1830024.

[178] J. Shen, T. Tang, and L.-L. Wang. *Spectral methods.* Vol. 41. Springer Series in Computational Mathematics. Algorithms, analysis and applications. Springer, Heidelberg, 2011, pp. xvi+470.

[179] J. Shen and J. Xu. "Convergence and error analysis for the scalar auxiliary variable (SAV) schemes to gradient flows". In: *SIAM J. Numer. Anal.* 56.5 (2018), pp. 2895–2912.

[180] J. Shen, J. Xu, and J. Yang. "A new class of efficient and robust energy stable schemes for gradient flows". In: *SIAM Rev.* 61.3 (2019), pp. 474–506.

[181] J. Shen, J. Xu, and J. Yang. "The scalar auxiliary variable (SAV) approach for gradient flows". In: *J. Comput. Phys.* 353 (2018), pp. 407–416.

[182] J. Simon. "Compact sets in the space $L^p(0,T;B)$". In: *Ann. Mat. Pura Appl. (4)* 146 (1987), pp. 65–96.

[183] Z. Songmu. "Asymptotic behavior of solution to the Cahn-Hillard equation". In: *Appl. Anal.* 23.3 (1986), pp. 165–184.

[184] A. Stevens. "The derivation of chemotaxis equations as limit dynamics of moderately interacting stochastic many-particle systems". In: *SIAM J. Appl. Math.* 61.1 (2000), pp. 183–212.

[185] R. Strehl, A. Sokolov, D. Kuzmin, and S. Turek. "A flux-corrected finite element method for chemotaxis problems". In: *Comput. Methods Appl. Math.* 10.2 (2010), pp. 219–232.

[186] R. Stupp et al. "Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial". In: *Lancet Oncol* 10.5 (2009), pp. 459–466.

[187] R. Stupp et al. "Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma". In: *N. England J. M.* 352.10 (2005), pp. 987–996.

[188] T. Suzuki. *Chemotaxis, reaction, network.* Mathematics for self-organization. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2018, pp. xii+316.

[189] I. Talhaoui, B. T. Matkarimov, T. Tchenio, D. O. Zharkov, and M. K. Saparbaev. "Aberrant base excision repair pathway of oxidatively damaged DNA: implications for degenerative diseases". In: *Free Radic. Biol. Med.* 107 (2017), pp. 266–277.

[190] L. Tang et al. "Computational modeling of 3D tumor growth and angiogenesis for chemotherapy evaluation". In: *PloS one* 9.1 (2014), e83962.

[191] U. Thiele and E. Knobloch. "Thin liquid films on a slightly inclined heated plate". In: *Physica D* 190.3 (Apr. 2004), pp. 213–248.

[192] V. Thomée. *Galerkin finite element methods for parabolic problems.* Vol. 25. Springer Series in Computational Mathematics. Springer-Verlag, Berlin, 1997, pp. x+302.

[193] G. Tierra and F. Guillén-González. "Numerical methods for solving the Cahn-Hilliard equation and its applicability to related energy-based models". In: *Arch. Comput. Methods Eng.* 22.2 (2015), pp. 269–289.

[194] A. Toma et al. “A validated mathematical model of tumour-immune interactions for glioblastoma”. In: *Curr. Med. Imaging* 9.2 (2013), pp. 145–153.

[195] L. N. Trefethen. *Spectral methods in MATLAB*. Vol. 10. Software, Environments, and Tools. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000, pp. xviii+165.

[196] S. Tremaine. “On the Origin of Irregular Structure in Saturn’s Rings”. In: *Astron. J.* 125.2 (Feb. 2003), pp. 894–901.

[197] J. M. Varah. “A lower bound for the smallest singular value of a matrix”. In: *Linear Algebra Appl.* 11 (1975), pp. 3–5.

[198] L. Wang and Y. Huang. “Error estimates for second-order SAV finite element method to phase field crystal model”. In: *Electron. Res. Arch.* 29.1 (2021), pp. 1735–1752.

[199] Y.-Y. Wang et al. “Adipose tissue and breast epithelial cells: A dangerous dynamic duo in breast cancer”. In: *Cancer Letters* 324.2 (2012), pp. 142–151.

[200] Z. Wang. “On chemotaxis models with cell population interactions”. In: *Math. Model. Nat. Phenom.* 5.3 (2010), pp. 173–190.

[201] Z. Wang and T. Hillen. “Classical solutions and pattern formation for a volume filling chemotaxis model”. In: *Chaos* 17.3 (2007), p. 037108.

[202] S. Ward. “Chemotaxis by the nematode Caenorhabditis elegans: identification of attractants and analysis of the response by use of mutants”. In: *Proc. Natl. Acad. Sci. USA* 70.3 (1973), pp. 817–821.

[203] S. M. Wise, J. S. Lowengrub, H. B. Frieboes, and V. Cristini. “Three-dimensional multispecies nonlinear tumor growth–I Model and numerical method”. eng. In: *J. Theor. Biol.* 253.3 (Aug. 2008), pp. 524–543.

[204] D. Wrzosek. “Volume filling effect in modelling chemotaxis”. In: *Math. Model. Nat. Phenom.* 5.1 (2010), pp. 123–147.

[205] R. Würth, A. Bajetto, J. K. Harrison, F. Barbieri, and T. Florio. “CXCL12 modulation of CXCR4 and CXCR7 activity in human glioblastoma stem-like cells and regulation of the tumor microenvironment”. In: *Front. Cell. Neurosci.* 8 (2014), p. 144.

[206] H. Yamaguchi, J. Wyckoff, and J. Condeelis. “Cell migration in tumors”. In: *Curr. Opin. Cell Biol.* 17.5 (2005), pp. 559–564.

Abstract

Nous étudions deux types de modèles couramment utilisés pour la représentation en temps et en espace des tumeurs : l'équation de Cahn-Hilliard pour les tissus vivants et le modèle de Keller-Segel. Les méthodes numériques que nous développons cherchent à représenter de manière précise et efficace ces équations tout en préservant leurs propriétés. Pour l'équation de Cahn-Hilliard, notre étude s'appuie sur une méthode de relaxation dont nous prouvons la convergence vers le modèle initial. Même si elles représentent mathématiquement des phénomènes physiques proches de ceux étudiés en dynamique des fluides, les équations utilisées pour les tissus vivants sont souvent différentes pour rendre compte du caractère actif des cellules. Les équations résultantes contiennent de nombreuses singularités et dégénérescences qui sont difficiles à analyser théoriquement et simuler numériquement de manière efficace. La méthode de relaxation a été introduite pour faciliter l'implémentation de nos schémas numériques ; nous proposons ainsi des schémas numériques éléments finis simples à adapter dans les codes pré-existants. Afin de préserver les propriétés des équations continues lors des simulations numériques, nous proposons des schémas numériques basés sur la Méthode de Variable Auxiliaire. L'adaptation de cette méthode pour les équations des tissus vivants n'ayant pas été réalisée, nous proposons dans cette thèse d'y remédier et d'étudier les propriétés analytiques de ces schémas numériques. Sur la base de ces travaux numériques, nous présentons l'étude de deux phénomènes biologiques. En collaboration avec des biologistes de l'Université de Nantes, nous étudions la compactification des sphéroïdes de glioblastome in-vitro en réponse à un médicament utilisé en chimiothérapie. Notre deuxième application s'intéresse à l'étude des effets physiques jouant un rôle dans l'émergence d'instabilités à la surface de certaines tumeurs invasives.

Keywords: Living tissues models, Numerical analysis, Degenerate Cahn-Hilliard equation, Keller-Segel model

---

Abstract

We study two classes of mathematical models currently used for the modeling in time and space of tumors: the Cahn-Hilliard equation for living tissues and the Keller-Segel model. The numerical methods we propose aim to represent these equations efficiently and accurately while preserving their properties. For the Cahn-Hilliard equation, our study is based on a relaxation method for which we prove the convergence to the original model. Even though the physical effects modeled by these equations are close to the ones studied in fluid dynamics, the equations used to model living tissues are different in order to represent the active behavior of cells. The resulting equations contain numerous singularities and degeneracies, which result in technical difficulties to analyze and simulate them efficiently. Our relaxation method has been introduced to facilitate the implementation of our numerical schemes. Hence, we propose numerical schemes that are easy to implement in already existing finite element software. In order to preserve the properties of the equations during numerical simulations, we design numerical schemes based on the Scalar Auxiliary Variable method. However, since this method has never been used in the context of models of living tissues, we study the analytical properties of our schemes. Based on these numerical works, we present two studies of biological phenomena. In collaboration with biologists from the Université de Nantes, we study the shrinking of in-vitro tumor aggregates of glioblastoma due to a certain chemotherapeutic drug. Our second study focuses on understanding the physical effects that play a role in the emergence of instabilities at the borders of certain invasive tumors. Therefore, this work aims at providing mathematical tools to biologists that give insights into underlying biological phenomena based on the Physics of cells and living matter.

Keywords: Living tissues models, Numerical analysis, Degenerate Cahn-Hilliard equation, Keller-Segel model

---

Laboratoire Jacques-Louis Lions
Sorbonne Université – Campus Pierre et Marie Curie – 4 place Jussieu – 75005 Paris – France