A Protocol for Model Validation and Causal Inference from Computer Simulation☆

Barbara Osimani[1,*], Roland Poellinger[1]

Abstract

The philosophical literature on modelling is increasingly vast, however clear formal analyses of computational modelling in systems biology are still lacking. We present a general, theoretical scheme which (i) visualizes the development and repeated refinement of a computer simulation, (ii) explicates the relation between different key concepts in modelling and simulation, and (iii) facilitates tracing the epistemological dynamics of model validation.

To illustrate and motivate our conceptual scheme, we analyse a case study, the discovery of the functional properties of a specific protein, E-cadherin, which seems to have a key role in metastatic processes by way of influencing cell growth and proliferation signalling.

To this end we distinguish two types of causal claims inferred from a computer simulation: (i) causal claims as plain combinations of basic rules (capturing the causal interplay of atomic behaviour) and (ii) causal claims on the level of emergent phenomena (tracing population dynamics). In formulating a protocol for model validation and causal inference, we show how, although such macro-level phenomena cannot be subjected to direct causal tests qua intervention (as, e.g., formulated in interventionist causal theories), they possibly suggest further manipulation tests at the basic micro-level. We thereby elucidate the micro-macro-level interaction in systems biology.

Keywords: Systems biology, micro-macro-level interaction, computer simulation, model validation, epistemology, emergent phenomena, causal inference

---

Contents

---

1. Introduction

Computational modelling in biology serves various epistemic goals by adopting different strategies. The main goal is to use models in order to simulate phenomena which cannot be observed *in vivo*: the integration of various pieces of knowledge serves the purpose of "observing", via simulation, the possible combined effect of various system components. Simulation is essential in that *emergent* behaviour is considered to result from the *iteration* of interactions among the individual components of the system, where the input conditions of iteration n depend on the interactions among the model components in iteration $n-1$ (and in a recursive way on all previous calculation steps; see below our discussion of calculation and measurement in Sec. 4.2).

2

By modelling and simulating the population-level behaviour resulting from causal interactions at the unit level, this methodology essentially differentiates itself from other available methods for causal inference (such as causal search techniques like those performed via causal discovery algorithms - see [1, 2], or standard statistical studies, such as randomised clinical trials, or observational cohort studies), which cannot model the micro-macro level interaction. We present here a case study in order to highlight such distinctive features with respect to more traditional methods of causal inference, and especially to advance a procedure for external validation of such techniques. In particular, we undertake a formal analysis of the epistemic warrant for such scientific procedures.

This account is motivated by several objections raised by philosophers regarding the scientific warrant of modelling and simulation for the purpose of scientific inference and discovery, especially in that this is at odds with the increasing use of such techniques as a complement and sometimes substitute to experiment, especially in the life sciences.[2]

This research is also urgent in view of the fact that much modeling and simulation research in biology has therapeutic purposes, e.g., the discovery and understanding of metastatic mechanisms in cancer. Also for this reason, we take as a case study, the discovery of the functional properties of a specific protein, E-cadherin, whose key role in metastatic processes has emerged from such simulation studies of cell growth and proliferation signalling.

In the second part of the paper we abstract from this case study and present a general theoretic scheme that relates the key concepts discussed in the philosophical literature on modelling and simulation in one and the same inferential framework (in particular, [3], [12], [14]). This is a diagrammatic map that represents the development process of the computer simulation:

1. The researchers start with (mathematically and/or logically) formalized assumptions about their object of interest (e.g., rules for an ABM, equations for an ODE system);
2. they translate these assumptions into a computer program and run simulations;
3. the simulation results are then compared to empirical observations and evaluated as to their quality in terms of reproduction of the observed behaviour of the target system;
4. the computer program is then repeatedly refined to generate results closer and closer to experimental findings (and biological background knowledge).

Nevertheless, in scientific practice, such a development process is rarely straightforward: first of all, revising the computer program might *increase* the discrepancy between (simulated) prediction and (physical) measurement. Yet, exactly by learning about such failures, scientists gain insights into the possible roots of such mismatches, guided by "how possibly" explanations of the repeated failures. If the revised model also fails to reproduce the target system, then the modeller has to find yet another explanation, while capitalising on the knowledge acquired through previous failures. In doing so, modelling failure helps refining the theory.

We present our protocol as a general, theoretical scheme which (i) visualizes the development and repeated refinement of a computer simulation, (ii) explicates the relation between different key concepts in modelling and simulation, and (iii) facilitates tracing the epistemological dynamics of model validation. In particular, we focus our analysis on the added epistemic value of modelling and

---

[2]Although the focus of this paper lies on computational modeling in systems biology, we build on a rich discussion of modeling and simulation strategies in the philosophy of science; e.g., [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14].

simulation techniques with respect to causal inference of micro-macro-level interactions. Also, we emphasize that causal knowledge at the microlevel can be drawn upon in order to put constraints on the possible structure underpinning the macro-level behaviour of the system under investigation.

Our paper is organized as follows: Sec. 2 provides an overview of the main modeling and simulation strategies in systems biology by focusing on the modal implications of the representation modes (discrete vs. continuous) and inference directions (top-down vs. bottom-up); Sec. 3 is dedicated to the presentation of a case study: the computer simulation of cell proliferation by Southgate, Walker and colleagues through three stages. Sec. 4 presents a formal account of the kind of inferential process illustrated by the case study, in view of providing a protocol for causal inference from computer simulation. In Sec. 5 we compare causal inference from computer simulation to other common techniques of causal inference available in the market, namely: causal discovery algorithms (see[1, 2]) and randomised controlled trials (RCTs). In Sec. 6 we embed the previous discussion in the debate around the epistemic nature of causal inference from computer simulation and present our concluding remarks.

2. Modelling and Simulation in Systems Biology

In systems biology, modelling is considered as a privileged instrument for knowledge integration and for the description/prediction/understanding of emergent behaviour. The main strategies available can be classified as a function of three fundamental categories: 1) mode of representation (continuum vs. discrete); 2) inference direction (bottom-up vs. top-down); 3) scale-adopted (single-vs. multiscale).

Representation Mode

1. The "continuum" strategy: this is generally represented by a system of (partial) differential equations, which individually represent the different forces or quantities of elements at stake, and whose joint effect is mathematically studied by means of the system itself.
2. The "discrete" or "rule based" strategy: this is for instance implemented by applying Agent Based Modelling to the behaviour of cells or molecules.

In both cases iteration allows for the "observation" of emergent behaviour (e.g. oscillations, or particular patterns of behaviour), that may (or may not) reproduce the investigated system. What distinguishes the two approaches is that the former focuses on quantitative relationships between the entities involved, such as the rate of reaction in catalytic processes; in the latter instead, the focus is rather on rules-based behaviour, such as the definition of the conditions under which a certain reaction is triggered rather than another one.

Inference Direction: model generation mode and modal character of the model

1. "Top-down" approach: starts with the observation of biological phenomena in the intact biological system (macro-level) and then constructs theories that would explain the observed behaviours, eventually uncovering the "underlying mechanisms" (meso-micro levels).
2. "bottom-up" approach: begins with the study of systems components in isolation and then integrates the behaviour of each component in order to predict the behaviour of the entire system (aggregating knowledge at the micro-level to "predict" behaviour at the macrolevel).

4

Bottom-up model are also considered to be "reductionist" to the extent that they rely on the assumption that higher-level phenomena can be analytically decomposed in modular, although holistically interacting, mechanisms.

The former proceeds via abduction (exploratory reasoning), whereas the latter by composition (integration of various pieces of consolidated knowledge). Both procedures are affected by uncertainty, in each case related to different kinds of underdetermination. In the former case, this refers to the various possible theoretical (and computational) models that can equally well reproduce the target phenomenon; in the latter, by the different possible ways in which the bits and pieces of local knowledge can be put together for global behaviour to be reproduced.

One could say that in exploring the space of possibilities, both approaches draw on consolidated knowledge; however, whereas top-down approaches draw on tacit knowledge concerning the possible formal rules that can aid in reproducing the phenomenon phenomenologically, bottom-down approaches are instead constrained in their composition process by local specific knowledge related to the items of the target system that they intend to reproduce. Top-down model are "toy-model" in that they refer to the target system holistically; whereas bottom-up model intend to refer to it on a one-to-one basis.

Top-down approaches generally constitute early-stage explorations of the space of theoretical possibilities. However, in the case study below we will see that toy models may be constructed also in a later stage of investigation, that is, after a series of "how possibly" explanations have been excluded by unsuccessful models, and empirical knowledge has put additional constraints on the causal structure underpinning the model.

Scales. Models can also be combined in order to span several spatial and/or temporal scales: integration of various models which span across levels of the system and represent temporal dynamics in the medium/long run is known as multiscale modelling. This sort of modeling approach is also termed "middle-out" approach in that it starts with an intermediate scale (usually the biological cell, as the basic unit of life) that is gradually expanded to include both smaller and larger spatial scales (spanning micro, meso, and macro scales; see, e.g., [15, p. 87]).

Regarding the inferential mode, the middle-out approach does not differ substantially from the bottom-up approach, in that it proceeds via composition of pieces of consolidated knowledge and hence is affected by the same kind of uncertainty. Scales of different magnitudes may also be temporal, so as to allow the model to incorporate subsystems working at different time scales; as for instance is the case of the cell cycle process (which occurs in the order of magnitude of hours), vs. receptor binding (seconds to minutes), vs. intercellular signalling (less than seconds).[3]

3. Case Study: Cell Proliferation Modelling

We present a case study based on the research conducted by Jenny Southgate, Dawn Walker, and colleagues on epithelial cell homeostasis and tissue growth, as well as wound healing phenomena

---

[3]For a taxonomic review of computational modelling approaches in biology see [16]. See also Marco Viceconti's discussion of integrative models with respect to their falsifiability in [14].

and malignant transformations [17, 18, 19, 20, 16, 21, 22, 20]. The investigated biological system is the set of mechanisms underpinning tissue growth and repair.

₁₃₀ We analyse here a specific line of investigation in their research program: modelling and simulation of the dynamics (and mechanisms) underpinning cell proliferation. This line of investigation can be broken down into three main stages, relating to the three different computational models developed by the research team:

1. *ABM.* The first model was a single-scale, discrete, bottom-up agent-based model, developed for the simulation of the population dynamics of cell proliferation,[4] which integrated in a unified system various local rules derived from experimental data ([17]). A comparison between the growth curves observed experimentally and those produced by the computer simulation revealed a significant discrepancy: two-way interaction in the experimental growth curve of the treatment setting, vs. a monotonic growth in the simulation curve. This discrepancy persuaded the scientists to revise the Agent Based Model by integrating it with a system of ordinary differential equations.

2. *Revised ABM + embedded system of ODE.* The second model was multiscale (cellular and subcellular level), multimode (discrete for the cellular and continuum for the subcellular level) model, which integrated the first model with a system of ordinary differential equations (ODE) representing a subcellular signalling pathway. This revised model investigated the hypothesis that the interaction effect observed experimentally could be explained by the interaction of the intracellular mechanism for cell proliferation (represented by the ODE), with the rules for proliferation dictated by cell:cell contacts (represented by the ABM) ([19]). However, in this case not only did the simulation not reflect the experimental curves, but not even the simulation "treatment" and "control" growth curve showed any significant difference.

3. *ABM, revised again, without ODE.* The third model eliminates the system of ODE introduced in the second one, which did not reproduce the expected behaviour, and adds a rule to the remaining ABM component. This "educated" toy model finally reproduces the growth curves observed experimentally.[5]

₁₅₅ In this iterative process, empirical knowledge puts constraints on the kinds of rules that the model is allowed to have (that is, progressively reduced the set of possible worlds), and the simulation results put further constraints on the kinds of *explanatory hypotheses for the observed mismatch* between target and source system (at the macro-level), which further reduced the theoretical space for the possible mechanisms at work (at the micro-level). Eventually, a model was found which reproduced the target phenomenon with sufficient accuracy. Our philosophical focus here is not on the vexed question of the ontological status of models with respect to their target (see, e.g., [23, 24, 25, 26, 7] or also [27, 28, 29]), nor on the kind of inferences that the similarity between

---

[4]Cell proliferation is an ambiguous term in that it refers both to the *individual* cell reproducing itself through mitosis (see below), as well to the growth of the cell*population*. Since the latter phenomenon is deterministically related to the former, this ambiguity comes with no serious epistemological consequence in this context. However, we will make a distinction between the two meanings clear whenever possible confusion may arise

[5]The agreement of the third model with the experimental data allows to learn, with reasonable confidence, that it is the temporal dimension of cell:cell contacts – rather than their spatial extension – that determines whether the involved cells will reproduce themselves or not. This constitutes the assumption on which a subsequent series of laboratory experiments is based and opens the way to more focused laboratory investigations, finally leading to the discovery of the complex context-sensitive regulatory role of E-cadherin in the homeostasis of cell proliferation [20].

target and source model allow (see, e.g., [30, 31, 32, 33, 34]), but rather on the process of discovery carried out through iterative model revision by means of finding possible explanations for
₁₆₅ the mismatches found along the process of refinement. Our main objective here is to identify the distinctive epistemic role and inferential process of modelling and simulation with respect to discovering the laws underpinning system-level behaviour and population dynamics. In particular, we aim to identify the specific inferential bonus provided by modelling and simulation when dealing with explaining dynamic phenomena that other standard techniques of causal inference (such as
₁₇₀ causal discovery algorithms and controlled experiments) fall short of explaining (see our comparative treatment in Sec. 5).

3.1. First Model: Bottom-Up ABM Modelling of Epithelial Cell Growth

The model is intended to simulate cell proliferation, in particular, it aims to verify whether by putting together the behavioural rules of the individual system components, established experimen-
₁₇₅ tally[6], the overall system dynamics can be reproduced computationally. The reason for simulating such behaviour is to verify whether assembling such local rules (causal relationships) is sufficient to reproduce ,*and therefore to explain on their basis*, the overall global system behaviour at the population level.

The left-side graph of Fig. 3 illustrates the causal structure being tested:

₁₈₀ 1. different environmental conditions (high/physiological vs. low calcium ions concentrations) determine different levels of the quality and functionality of the protein E-cadherin present in cell culture;

2. on its turn, presence vs. absence of E-cadherin, determines the type of cell:cell contact: stable vs. transient;

₁₈₅ 3. The number of contacts per cells at any given time determines proliferation (< 4), or inhibition of proliferation (≥ 4);

4. migration is inhibited by stable contacts.[7]

The computational model was matched to a lab experiment conducted on normal human urothelial cells (NHU), comparing cell growth in physiological (2.0mM) vs. low (0.09mM) calcium concen-
₁₉₀ tration medium. The comparison of the growth curves in the two settings showed similar patterns but also a significant discrepancy (compare the plots in Fig. 2 against the benchmark provided by Fig. C.11 in Appendix C):

1. In agreement with what was "predicted" by the computational model, cell growth was *higher in the low calcium medium*, where the phenomenon of contact inhibition is significantly weaker
₁₉₅ (see Fig. 1).

2. In contrast to the curves representing cell growth in the computational simulation, where the lag phases in low and physiological calcium concentrations were similar (see Fig. 2 left), the results from the experiments showed that *the lag phase for the low calcium culture system*

---

[6]Southgate and colleagues emphasize that their model distinguishes itself from previous models proposed for the same phenomenon whose "rules are selected to produce the desired outcome, and are not linked to the underlying biological mechanisms" or whose scope covered "limited range of cell behaviours". The rules should represent "shortcuts" to *actual* controlling mechanisms. Hence, the model is not "phenomenological".

[7]Details about the model can be found in the Appendix B. The rules dictating cell behaviour can be found in Appendix A.

was much longer than for the physiological calcium medium (see Fig. 2 right). Or else, that *in low density populations, growth was higher in physiological than in low calcium concentration medium*; in contrast to the hypothesis that, by inducing stable contacts, E-cadherin (present in high calcium concentrations) would inhibit proliferation.
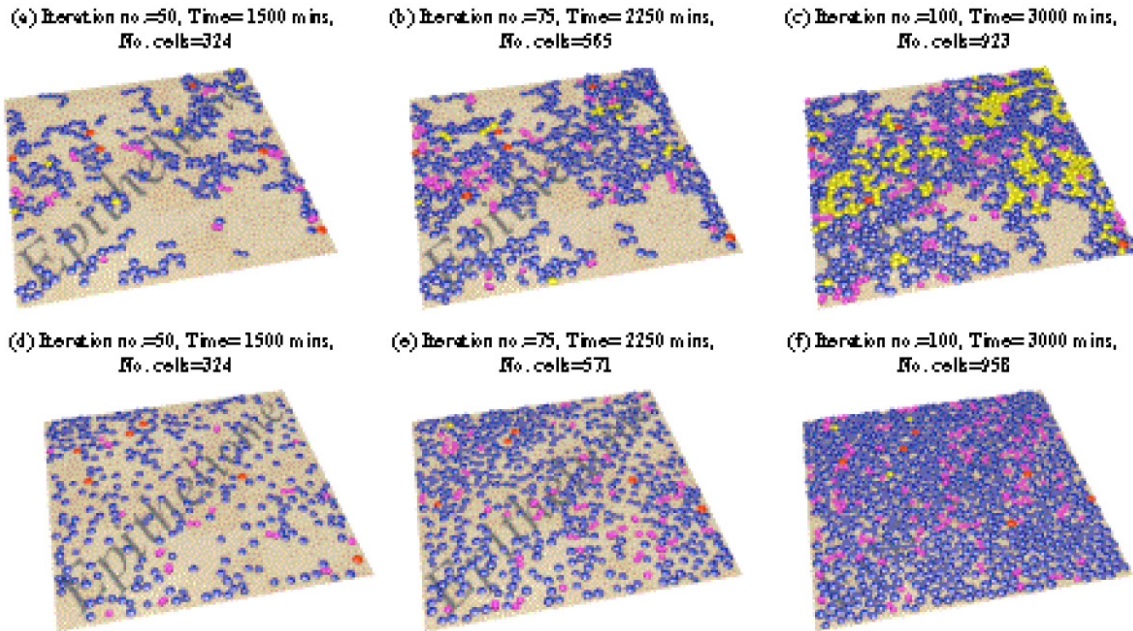

Figure 1: Pattern of simulated culture growth in physiological calcium (a–c) and low calcium (d–f). Snapshots correspond to iteration 50 (a, d), iteration 75 (b, e) and iteration 100 (c, f). Red cells represent stem cells, blue cells are transit-amplifying (TA), pink cells are mitotic, and yellow cells are quiescent. The growth pattern in low and medium density populations (figures a and b in comparison with d and e) is relatively similar, because contact inhibition is less likely to have an influence in sparse populations. Instead as long as the population density increases the simulation shows diverging patterns. In the subconfluence phase (c and f) contact inhibition has a stronger influence in the physiological calcium environment (c), where contacts tend to be more stable, and this is reflected in decreased migration and the formation of cells clusters. Figure reprinted from [17] with permission from Elsevier.

Hence, Walker and colleagues looked for an explanation for such discrepancy, and searched for possible hints from the empirical literature: this suggested that autocrine signalling (a sort of cell self-signalling) induces growth transcription factors, such as Erk ([35]).

They hypothesised that, under the threshold for proliferation inhibition (≥ 4 cell:cell contacts), *stability of the contact itself might indeed induce proliferation through cell:cell signalling*. Thus the effect of E-cadherine would be "bi-directional" and would induce:

1. both *inhibition of proliferation*; by promoting stable contacts and thereby increasing the probability that at any given time any given cell has ≥ 4 contacts – this was the main hypothesis underpinning the original model about cell growth rules.
2. and *promotion of proliferation*; through cell:cell signalling – this constituted the revised hypothesis.

8

The revised hypothesis does not replace the original one, but rather specifies it further, so as to possibly uncover the origin of the interaction between population density and low vs. physiological calcium conditions observed experimentally.
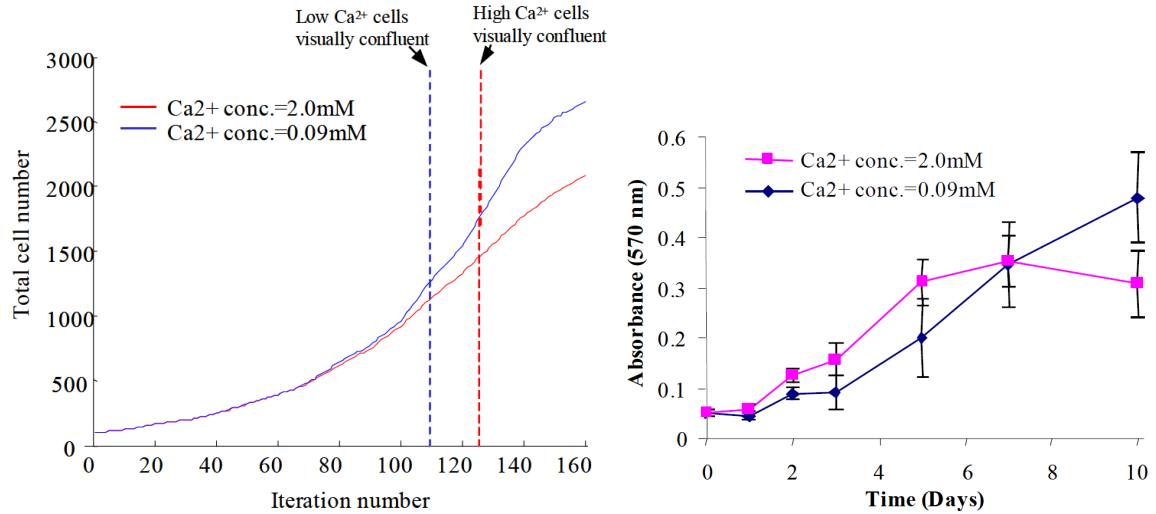

Figure 2: Left: Growth curves produced by computational model (linear scale). Right: Results of MTT assay (error bars represent standard deviation of 6 samples). Figures reprinted from [17] with permission from Elsevier.
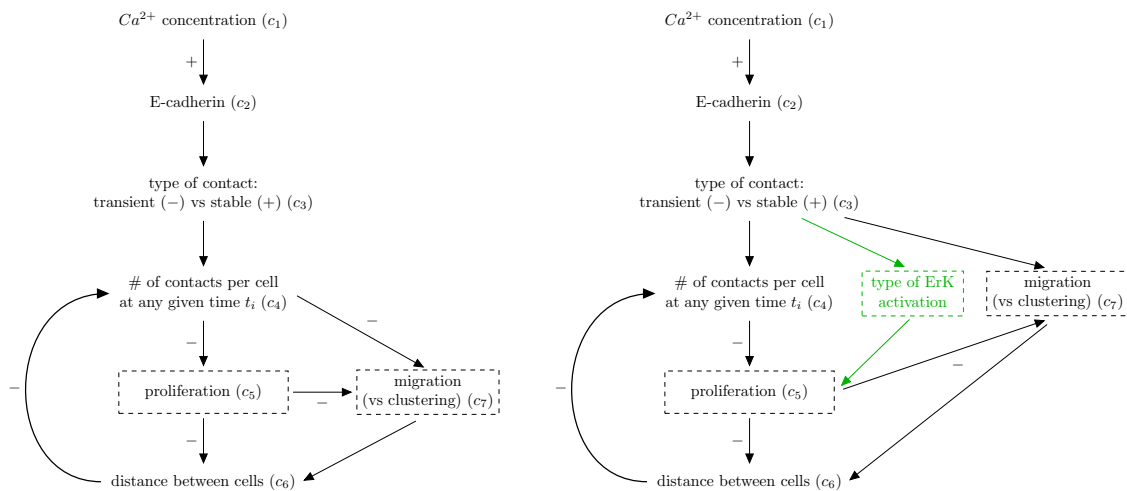
Figure 3: Causal structure underpinning the first ABM of epidermal cell growth and proliferation (left) and extended causal structure where the type of ErK activation mediates type of contact and proliferation (right), all causal paths in the second model are represented by the same (or slightly modified) ABM rules of the first model, with exception of the causal path from "type of contact" through "type of Erk activation" to "proliferation", that is the intracellular signalling pathway, which is instead represented by a system of ODE standing for biological rates of reactions observed in laboratory experiments; see [17].

9

3.2. Second Model: Integration of the First Agent-Based Model and an ODE System into a Multiscale Model

In order to explore the "augmented" hypothesis, the scientists devised a second model where
also intracellular signalling pathways could play a role in cell growth regulation. This was done
by integrating the agent based model with a system of ordinary differential equations aimed at
reproducing the (supposedly growth promoting) intracellular signalling pathway, and by iterating
the ABM component and the ODE system in alternation.[8] Figure 3 compares the two hypothetical
causal structures underpinning the first and the second model respectively.

The second model [19] was meant to explore the two counteracting effects of E-cadherin: 1)
inhibition of proliferation through promotion of stable contacts, and 2) promotion of proliferation
through sustained activation of Erk (exactly mediated by stable contacts) at low vs. high density
population.[9] It was applied to two settings:

(a) *individual* cell:cell contact (for different rates of formation of the contact);

(b) a growing *population* of cells (in simulated low vs. physiological calcium conditions).

The first setting was meant to compare the effect of stable vs. transient cell:cell contact on Erk
activation levels for one-to-one intercellular contacts. The different patterns are shown in Fig. 4
below.

In both cases the models are meant to reproduce the high/physiological calcium conditions vs.
low calcium environments (with scarcity of E-cadherin mediated contacts). Whereas the individual
cell:cell contact model (a) delivered results congruent with expectations (and with empirical data);
the population level model (b) gave puzzling results, since not only did the model not reproduce
the growth curve that it was expected to deliver according to the hypothesized causal structure,
but also the growth curves in the two conditions ("low vs. "high calcium") did not differ from
each other. This was puzzling at first, given that in both (a) and (b) the signalling pathways were
modelled by the same set of ODE.

However, the initial conditions for model (a) and (b) where not the same.

In model (a), the initial conditions for three different settings were (see: 4)[10]:

---

[8]The set of equations of the ODE model representing Erk activation (i.e., Erk diphosphorylation: Erk-PP) through
the EGFR signalling pathway, takes as initial value the number of receptors and ligands binded between two adjacent
cells. The total number of activated receptors will depend on the number n of contacts and their areas, which is
fixed in the low calcium setting, and *constantly grows* in the high calcium one). Each cell has a set of equations 1–5
associated with each cell:cell contact; these are then summed to provide the total number of activated receptors per
cell, which in turn provides the relevant quantity for the subsequent equations in the system ([19]). The last equation
of the ODE model (eq. 29 in [19]) represents the level of Erk activation for that cell in that stage of the simulation.
At the end of the ODE iteration, the updated profile of Erk-PP is returned to the memory of each agent of the ABM.
Hence, the Erk activation type determines for each cell whether the cell divides and proliferates or not; this in turn
will change the density (and thereby the topology) of the population of cells at the new start of iteration, which, in
turn, will determine which cells will have new contacts and whether these will determine proliferation (if these have
been stable enough through iterations) or not (if they have been not stable, or if the number of contacts reached by
the cells is ≥ 4 and therefore the "contact inhibition" rule trumps the signaling pathway effect). The constants for
rates of reactions are all derived from experimental studies ([36, 37, 38, 39, 40, 41, 42]).

[9]In the new model however, the previous ABM component is not only integrated with an ODE system, but also
modified – see Appendix D.

[10]Experimental evidence shows that the temporal characteristics of ERK activation (through the EGFR pathway)
influence cell fate: moderate but sustained ERK activation lead to cell progression (and proliferation), whereas
transient activation does not ([36, 37], see Fig. 4 below and related explanation).

1. moderate but sustained rate of contact formation;[11]

245 2. same initial formation pattern as for case 1, but with abrupt break of contact after 4 hours;[12]

3. "instantaneously" formed contact, but quickly decreasing its strength already after 20 minutes.[13]

To these three different conditions correspond different patterns of Erk diphosphorilation (that is, of activation): only in the first case the cell reproduces itself.[14] Case 1 and 3 are supported by
250 experimental evidence ([19, 38]).

In model (b) the two initial conditions were:

1. linearly increase the area of cell:cell contact for each pair of cells entered into contact at each iteration until max length of 1,5 times the cell radius (and $5\mu m$ height of the contact area);

2. do not increase contact area through iterations.

255 As can be noted, the difference in the duration of the contact is not encoded in this set of initial conditions. This revealed to be crucial to explain the mismatch between the individual level and the population level simulation, and also in devising the third model.

By observing the results of the two simulations, and considering the difference of initial conditions and related outputs, the scientists could infer that Erk activation depends on the temporal
260 characteristics of contact formation only, *irrespective of its spatial extent*, and that it is through this dimension that the interaction effect in population growth dynamics is produced at the macro-level.

Hence the discrepancy is explained by uncovering a hidden, wrong, assumption, which is implicit in modelling the two conditions. Namely, the assumption that Erk activation would depend on the spatial extent of the contact; whereas, *as the one:one simulation showed*, this depended on the
265 continuity of the contact in time only. The initial conditions for the individual level simulation fix the rate of growth in time exogenously, and therefore the ODE outputs three different curves depending on these initial conditions.[15] Instead, in the population level model, the two settings only differ for the growth of the *spatial* dimension of the contact in time, vs. no growth.

Since in the individual cell:cell contact simulation the different initial conditions give rise to
270 different simulation outputs, whereas this does not happen in the population-level simulation, the explanatory hypothesis is that the ODE system is sensitive to temporal duration of the contacts only, and not to their spatial length (or only indirectly so). A stable (even if non E-cadherin mediated) contact, may show exactly the pattern of Erk activation of type 1 in Fig. 4 if it lasts long enough.

275 The mismatch between the expected simulation patterns at individual vs. population level on one side, and the acquired knowledge that at the individual level the different patterns of Erk activation are determined by the temporal length of the intercellular contact only (consolidated by correspondence of three temporal patterns and Erk activation curves with independent experimental data, cf. [19, 38]), finally led to a third (single-scale) ABM model.

---

[11]initial contact of $2\mu m$ that increases linearly to $23\mu m$ over 6 hours.

[12]when it has reached a contact length of $16\mu m$.

[13]when it has barely reached $12\mu m$

[14]Large, instantly formed contacts induce a large peak in Erk activation at ca. 20 min. after the initial contact, but then a rapid decline to very low levels follows, even though the contact persists. Instead a contact which forms gradually but persists in time produces a monotonic, although moderate, increase in Erk activation. If the contact is lost (case 2), then the Erk-PP signal quickly goes down to 0.

[15]For an overview over how the modelling assumptions were tested, see Appendix E.
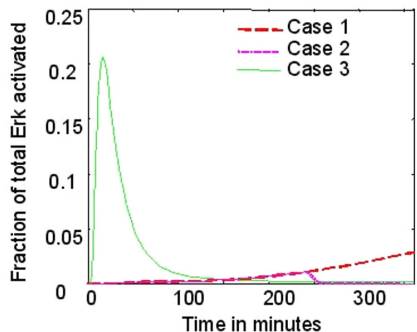
Figure 4: Erk activation patterns (measured by normalized proportions of Erk-PP) for a single cell:cell contact. The dashed, monotonically increasing line (case 1) represents the case where two cells make an initial contact of 2μm that increases linearly to 23μm over 6 hours. This type of behaviour has indeed been experimentally observed in (18). The continuous dark line (case 2) shows the same initial formation pattern as for case 1, but then has an abrupt break of contact after 4 hours (when it has reached a contact length of 16μm). The continuous light line (case 3) shows a strong initial signal which is "instantaneously" formed, but quickly decreases its strength already after 20 min., when it has barely reached 12μm. Case 3 is typical of those contacts observed (by the authors) using time-lapse microscopy between migratory cells, maintained in subphysiological medium (Ca^{2+} concentration: 0.09mM). Figure reprinted from [19] under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0).

₂₈₀ *3.3. Third Model: Simplified Educated-Phenomenological Model*

The third model appears in yet another study ([21]), which black-boxes the effect of the signalling pathway(s), and substitutes the mathematical modelling of the signalling cascade (the ODE system embedded in the ABM), with a simple agent-based rule, forcing the cell to increment its cycle step and reproduce itself whenever it has maintained a contact for ≥ 8 hours. Hence this model consists ₂₈₅ of the ABM component of the second model only – updated with the additional rule that the cell should proceed in the cell cycle if it has maintained at least one contact, *whether "transient" in the sense of non-calcium mediated, hence with fixed dimension, or stable, i.e., calcium-mediated, hence growing in dimension,* for at least 8 hours[16] ([21]). This means that the only parameter which is allowed to influence proliferation is the temporal length of the contact, whereas the spatial ₂₉₀ extension of the contact plays no role.

The model can be considered to be an "educated"-phenomenological model in that the additional rule aims to reproduce the expected cell population growth curve experimentally observed in the target system, with no assumption as to the biological mechanisms underpinning it. However, it serves the purpose of testing the hypothesis that the temporal length of contacts can account for ₂₉₅ the growth curves observed experimentally.

In this case, the growth curves of experimental and simulated data finally matched each other. This means that whatever the signaling pathway at play, and whatever its intrinsic cascade, its effect on population growth may be (fully) explained by the temporal length of the contact. Is this conclusion warranted? If so, how and what are its implications?[17]

---

[16] Please note that in the second modelling attempt, cell proliferation (self-reproduction) was triggered in the individual cell:cell contact model (a), by a stable contact lasting more than 6 hours.

[17] Keeping fixed the knowledge acquired through the various simulations – that is, the positive role of E-cadherin and of the temporal length of intercellular contacts in modulating cell growth – while questioning the role of the EGFR-Erk pathway in cell proliferation, the authors undertook then a series of experimental investigations on

4. Towards a Protocol for Causal Inference from Computer Simulation

4.1. From Formal Model to Stable Code

When Margaret Morrison discusses the epistemic value of computer simulations ([12]), she emphasizes the twofold role of models in our attempt at grasping reality. She states that "[...] models are experimental tools capable of functioning as measuring instruments, but they are also objects of inquiry in cases where they, rather than physical systems, are being directly investigated and manipulated. This suggests an analogy between models and experiment, and "[...] it is this link that figures prominently in establishing the experimental character of simulations." (see [12, pp. 228–229]). Of course, when Morrison speaks of *models as objects of inquiry*, she does not mean to conflate the model and what is being modeled. She does not claim that manipulating the model is all we need to do to learn about the modeled object or system. Quite the opposite: Morrison identifies the methodological and epistemological characteristics of physical experimentation and applies this characterization to model manipulation, thereby providing an interpretation of models beyond the classical understanding as descriptive tools. In our conceptual framework we build on Morrison's vocabulary and aim to not only (i) tie real physical systems to their formal stand-ins, but also (ii) link such formal models to higher-level phenomena emerging from computer simulations. Our conceptual scheme, which spans all these epistemological/methodological levels, offers a unified way of tracing scientific inference – "upwards" from basic causal assumptions to complex causal claims, and "downwards" from emergent higher-level patterns to causal relations underlying the target system. Such higher-level phenomena are of special interest to us in this paper, since they facilitate causal inference in a special way: not only are emergent phenomena grounded in a simulation's basic rules, but they truly only arise in the dynamic evolution of the simulated system. Nevertheless, higher-level behavior evades both (i) Morrison's characerization of model manipulation as experiment and (ii) the concept of intervention as defining causal relations (as, e.g., formulated in interventionist causal theories). Still, although the relation between the simulation model and higher-level observations can be very opaque, emergent phenomena are nevertheless *generated* by aggregation/interaction of basic rules. So, when running a simulation leads to the detection of emergent patterns, the simulation acts as a *guide to causal understanding*. In particular, perceived macro-level phenomena will possibly suggest further manipulation tests at the basic micro-level. In this way, emergent phenomena help in tracing the interplay of the target system's mechanisms (and sub-mechanisms): i.e. the micro-macro-level interaction.

In the following, we abstract from the case study above and introduce a general theoretic scheme to visually relate the key concepts discussed in the philosophical literature on modelling and simulation. This scheme reflects the development process of the computer simulation: The researchers start with formal scientific assumptions about their object of interest (commonly mechanistic knowledge in mathematical form) and translate those assumptions into a computer program. Ideally, this computer program is then repeatedly refined to generate results closer and closer to experimental findings (and biological background knowledge).

Figure 5 traces the process from the initial formal model of the target system to its simulation in a repeatedly revised code base. The computer program can be understood as a set of structured

---

alternative E-cadherin related signalling pathways in normal human urothelial cells, finally coming to the conclusion that this protein has a complex, context-sensitive behaviour ([20]).

equations or rules $\mathcal{R}$, called in a predetermined order to update the model's parameters at each stage of the simulation. Since $\mathcal{R}$ is revised if simulation and empirical evidence deviate unacceptably, index k indicates the "version number" of the program: The simulation model is implemented as $\mathcal{R}^1$ and revised until the stable version k^* satisfies the researcher.

In scientific practice, such a development process is rarely straightforward: For example, revising the computer program might unexpectedly *increase* the discrepancy between (simulated) prediction and (physical) measurement and thus reveal gaps in the initial assumptions. Yet, learning about such theoretical gaps might then motivate the next revision of the computer program, and – in doing so – help in refining the theory. Our graphical scheme relates the computer program and the theoretical background by marking constraints for code revisions, grounded empirically and on abductive bases. As the case study presented above illustrates, revisions are made on the bases of "how possibly explanations" for the observed results constrained by empirical findings and consolidated background knowledge. On their turn, new simulations results further constrain the space of possible explanations.

Our general theoretic scheme can not only be utilized (i) as a blueprint for tracking the development process of the computer simulation, but also (ii) as a basis for justifying causal claims drawn from such computer simulations. In Secs. 4.2–4.5, we build on this scheme in order to outline the basis for a general protocol intended to support justified causal inference from computer simulations. We proceed as follows:

Before we can start to compare quantities measured in simulation and in empirical experiment, we define the concept of *measurement by simulation* (Sec. 4.2). In Sec. 4.3, we present potential sources of discrepancy between experiment and simulation and discuss ways of revising the computer program while preserving consistency of the codebase. We then introduce the task of model validation as a question of accepting a model once it reaches the desired robustness, and we sketch this robustness condition in terms of multi-dimensional accuracy (Sec. 4.4). Our aim in Sec. 4.5 is to go beyond reconstructing phenomenologically accurate prediction, and facilitate understanding of the causal dynamics at work. To this end we formulate an anchoring condition a (sufficiently) robust simulation must satisfy (across code revisions), in order to warrant valid causal inference (i) from compound causal information, and (ii) from emergent phenomena. By explicating an epistemological concept of emergence we finally have all the ingredients for a complete protocol for causal inference from computer simulation. We start by distinguishing calculation and measurement.

4.2. Measurement by Simulation

When the computer program is run, it simulates the evolution of the target system by iteratively updating the model's parameters in temporal increments of Δt intervals over n stages. To capture as much detail as possible, Δt should be as small as possible without unnecessarily increasing computing load or processing time. At each stage s_i $(1 < i < n)$ of the simulation, the parameter values are calculated from the results of the previous stage s_{i-1}. For instance, in our case study, each cell's properties such as size, location, cell cycle phase etc. are calculated given the parameter values calculated in the previous calculation step. This calculation cycle, generating interpretable data over n stages, can then be called measurement, even if the computer experiment lacks the
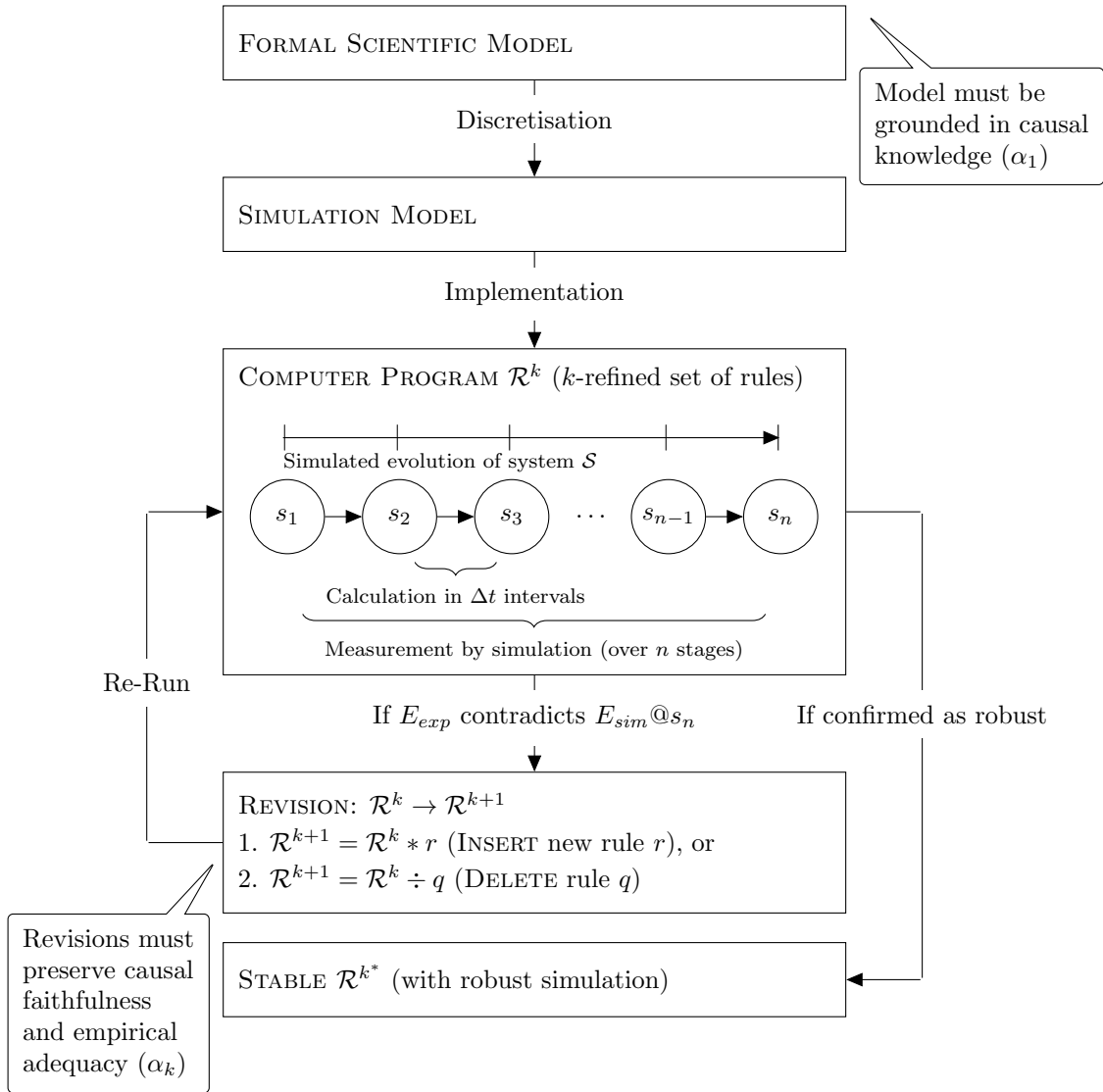
Figure 5: Tracing the process from model to robust simulation in runs of the computer program $\mathcal{R}^{k^*}$ (cf. Morrison 2015): A scientific model is transferred into a simulation model, implemented as a computer program, and revised until the simulation yields robust results.

material link to the target system.[18] For example, the ODE system in the second model of our case study was used to virtually *measure* Erk activation patterns for a single cell:cell contact (as

---

[18]This characterization follows Morrison's discussion of calculation and measurement in simulation and scientific practice in general, see [12, chpt. 6, pp. 199ff.].

illustrated in Fig. 4).

4.3. Verification, Validation, Revision

If empirical data is available, the physical, experimental evidence, abbreviated as E_{exp}, can be compared to the simulated quantity, abbreviated as E_{sim} (encoding the measurement result of n calculation steps). If the quantities deviate unacceptably (for a predefined minimum level of accuracy) the computer simulation is deemed invalidated, and $\mathcal{R}$ must be revised in order to improve the simulation.

The modeler is presented with the oftentimes difficult task of judging what such a discrepancy might be due to (see also our discussion of our case study's modelling assumptions in Sec. E). Assuming that the experimental, empirical results are correct, one of the following situations might obtain:[19]

1. One of the idealizations (deliberately chosen or as constraint of the programming framework) applied to formulate the mathematical representation of the mechanistic knowledge has a severe effect.
2. The numerical solution of the mathematical form used to capture the mechanistic knowledge is inaccurate.

The above issues typically regard the verification phase of a computational model: Code verification ensures that the program runs without errors, and solution verification ensures that any deviations of expected outcome values are not due to how the program's functions are implemented. Discrepancies due to cases (1) or (2) can only be remedied by revisiting the simulation model or basic implementation decisions.

When we turn to validation issues, the discrepancy between simulated and experimental measurement might be due to one of the following cases:

3. The initial parameter values (at stage s_1) are inaccurate.
4. The mechanistic knowledge the model is based on is wrong or incomplete.
5. The mathematical form used to capture the mechanistic knowledge is inadequate.

Cases 3–5 refer to information encoded in elements of $\mathcal{R}$, i.e., structural equations containing constants and variables. Discrepancies due to this second list of cases reflect problematic assumptions in the modeler's point of departure, i.e., her initial formal representation of the object of inquiry in the mathematical model. Emily Parke ([3]) makes out discrepancies of different strength in her discussion of the supposed epistemic privilege of real experiments over computer simulations. She first distinguishes different degrees of surprise, as a researcher's reaction to unexpected observations: i) The researcher might find the observations unexpected, but in general agreement with background assumptions and theoretical considerations, or ii) she might be puzzled to such a degree that she is forced to reevaluate or completely overturn her theoretical assumptions.

Since this distinction is relative to the generality of the background assumptions, Parke shifts her focus to two relevant sources of empirical surprise *within the object under investigation itself*:

1. the investigated system evolves unexpectedly, or

---

[19]We are grateful to Marco Viceconti for urging us to be explicit about the following range of sources of discrepancy.

2. the evolution of the investigated system reveals *hidden mechanisms* or *hidden causal factors*, thereby exposing implicit basic assumptions or detecting gaps in the simulation model.

In our case study, it is precisely such a gap in the simulation model that prompts further refinement, and "hidden" precisely refers to the role played by micro-level forces and entities in the output observed at the macro-level. Hence, the disagreement between simulation and empirical experiment which Walker et al. observe in their first model can be described as an instance of case 4 above. By looking for possible explanation of such a discrepancy and especially by looking at the form of growth curve of the cells population in the experimental setting, they *must* assume the presence of a factor interacting with the contact-inhibition rule. Then, by looking for possible candidates, they learn from cell biology literature about the possible positive effect on proliferation, caused by intracellular signaling pathways activated by cell:cell contact. Hence, they include a signalling pathway in the model in the form of a set of equations derived from empirical data, and reiterate the simulation. In doing so, the scientists add fine-grained detail to their assumptions about what is of relevance for a mechanistic understanding of cell proliferation.[20] In principle, revisions of $\mathcal{R}$ can be translated into revisions of the set of theoretical assumptions about the target system.

Our abstract scheme in Fig. 5 presents the revision step from one version of the program to the next, i.e., from $\mathcal{R}^k$ to $\mathcal{R}^{k+1}$, in the syntax of the belief revision formalism (as, e.g., developed in [Gärdenfors 1988]):[21]

1. Insertion: $\mathcal{R}^{k+1} = \mathcal{R}^k * r$. The rule r is added to $\mathcal{R}$ while maintaining consistency, i.e., after any rule in conflict with r is removed from $\mathcal{R}$. In belief revision terminology: $\mathcal{R}^k$ is *revised* by r.
2. Deletion: $\mathcal{R}^{k+1} = \mathcal{R}^k \div q$. R^k is *contracted* by rule q, which means that any rules implying q together with any rules implied by q are removed from $\mathcal{R}^k$ (i.e., all traces of q are eliminated). This operation might be realized in more than one way, since contraction does not have a unique solution in general. Causal information, e.g. from empirical evidence, might be of help in disambiguating possible options (see also the remarks on condition α below in Sec. 4.5).

Possible strategies for model revisions (through rule insertion and deletion as well as combinations thereof) are illustrated in the following sections. By revising the computer program, the modeler hopes to approach a stable code-base step-by-step. Is there a way for her to judge the quality of the simulation, though, and to validate the model conclusively?

4.4. Accuracy and Robustness

The goal of the modeler is to create a reliable and accurate simulation of the target system. Accuracy and reliability are verified in multiple dimensions:

---

[20]We find it interesting to note how the modification of the ABM component in the third model remedies the observed discrepancy: The fact that it is a (phenomenologically motivated) local refinement of the (otherwise mechanistically motivated) ABM rules, which suffices to correct the higher-level measurement, seems to suggest that the ABM is overall adequate, yet underspecified (thus pointing to case 4 more than to case 5 among the sources of discrepancy listed above).

[21]This formalism provides ready-made operations for consistently updating sets of interdependent sentences (in our case: rules or lines of code). However, in principle, of course, any other formalism will do to formally capture the operations of rule insertion and deletion.

1. *Point-wise comparison* of simulated outcome and empirical measurement might be deceptive, because agreement might merely be due to the simulation model being fine-tuned in calibration, i.e., due to the model crucially depending on its very input parameters (often referred to as "overfitting problem"); see our discussion of the comparison of the first model's predictions with experimental data (illustrated in Fig. 2).

2. If the modeler wants to confirm her understanding of the mechanistic structure though, she will have to check whether the simulation generates robust outcomes even under *variation of parameters* (within empirically testable ranges)[22];

3. Once the simulation reaches the desired level of accuracy even under such variation of parameters, the modeler may – in a further step – subject her model to *mechanistic intervention* and check whether the simulated evolution of the target system still corresponds to experimental evidence from the target system (again, with desired accuracy), for example, when the researcher blocks a mechanism in the experimental set-up and simultaneously disables the supposed corresponding rule in the model. In our case study, locally embedding the ODE system into the agent-based model can be understood as a starting point for such mechanistic tests.

The process of scrutinizing the model by varying the model's parameters in static comparison, and by intervening on the model's mechanisms, (revision), eventually leads to a stable code-base $\mathcal{R}^{k^*}$ that *robustly* generates phenomena corresponding to empirical experiments with desired accuracy, thereby confirming the model – or, to be precise, the model's adequacy (and thus its applicability).

The link between simulation and target system can then be expressed in terms of *robustness*: The concept of robustness of a simulation captures a simulation's accuracy under parameter change *and* mechanistic intervention – robustness can thus be understood as "multi-dimensional accuracy". If the simulation returns results above a certain empirically motivated degree of accuracy, the researcher is justified in postulating an *analogous relation* between simulation model and target system. Optimizing this measure of robustness will thus *confirm* the simulation model as being *adequate* for making inferences about the target system.[23]

Example. Fig. 6 illustrates the concept of robustness by tracking the evolution of an unspecified example system and its simulation over time for three different initial states. Solid lines indicate experimental evidence, whereas dashed lines represent the simulated evolution of the target system for the same initial states. The simulation is obviously well-calibrated for initial state (a) (and for a short initial time segment across initial states), but parameter changes reveal its non-robustness: The simulation increasingly deviates over time from experimental results for initial states (b) and (c).

In our case study, the researchers want to achieve accuracy in two important respects: (i) not only for individual cell:cell contact (for which the second model seemed to be well-suited), but also at

---

[22]see our discussion of test 1 (sensitivity analysis) in the case study in Appendix E

[23]In terms of Bayesian confirmation, increasing the robustness and thereby increasing the belief in *analogous evolution* of simulated (*SimS*) and target system (*TarS*), formalized as '*SimS* $\approx$ *TarS*', will boost the researcher's belief in the adequacy of the simulation model (*SM*), on which the simulation is based:

$$P(SM \text{ adequate} \,|\, \overbrace{SimS \approx TarS}^{\text{Robustness}}) > P(SM \text{ adequate}).$$
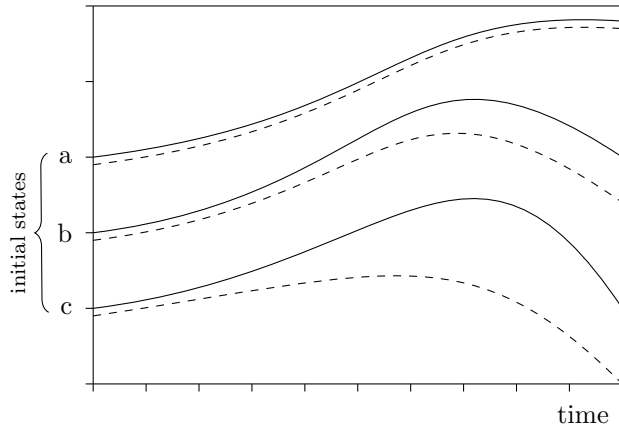
Figure 6: Robustness by example: Solid lines indicate experimental evidence, dashed lines represent the simulated evolution of the target system. The simulation model is well-calibrated for initial states (a) and deviates with decreasing accuracy for initial states (b) and (c). The simulation thus reveals *non-robustness* under parameter changes.

the population level (for which the second model failed); and (ii) not only in quantitative comparison between simulation and experiment (as by matching statistics), but also in qualitative comparison (as by comparison of structural macro-level dynamics, see also our discussion of inference from emergent phenomena below in Sec. 4.5).

Now, where maximizing accuracy under parameter change may increase phenomenological agreement, maximizing accuracy under mechanistic intervention is the first step towards a mechanistic understanding of the underlying causal relations, since it leverages causal control for calibration. In the following we want to tie accurate prediction to basic causal knowledge by explicitly considering anchoring conditions for justified causal inference from the aggregate level.

4.5. Causal Inference

The philosophical skeptic might argue that simulations as virtual, artificial environments are detached from reality – they follow their source code and will behave like programmed, not like reality, which renders such computer simulations unusable for scientific – not to mention *causal* – inference. This claim is pointedly summarized by Diez Roux ([43, p. 101]) in her discussion of the distinction between observation-based and simulation-based causal inference in epidemiology:

> [...] there is a fundamental distinction between causal inference based on observations (as in traditional epidemiology) and causal inference based on simulation modeling. The traditional tools of epidemiology are used to extract (hopefully) reasonable conclusions from necessarily partial and incomplete (often messy) observations of the real world. [...] In contrast, when we use the tools of complex systems, we create a virtual world (based on prior knowledge or intuition) and then explore hypotheses about causes under

19

the assumptions encoded in this virtual world. In the simulation model, we cannot directly determine whether X causes Y in the real world (because the world in which we ₅₂₀ are working is of our own creation); we can only explore the plausible implications of changing X on levels of Y under the conditions encoded in the model. In the real world, we have fact (what we observe) and we try to infer the counterfactual condition (what we would have observed if the treatment had been different). In the simulated world, everything is counterfactual in the sense that the world and all possible scenarios are ₅₂₅ artificially created by the scientist.

This topic has recently seen a surge of publications in the philosophy of science with some authors regarding computational modeling as simply one variant of modeling as such (e.g., [10]), while others emphasize as a special feature of computer simulations the possibility to experiment with models as objects of inquiry (e.g., [4] and [12]). In line with the idea of computational models as test objects ₅₃₀ we described the code-base of the simulation as constantly changing with rules inserted or deleted in attempts at increasing the accuracy of the program (see Sec. 4). In the following we focus on the question how causal inference from iteratively revised simulated systems can be justified, if at all. We proceed in a bottom-up way: In order to guarantee the legitimacy of causal inference from computer simulations, the simulation model must be built on a solid knowledge base. The empirical ₅₃₅ adequacy of this foundation permits causal inference from the computer simulation with respect to two distinct dimensions of aggregation: i) inference from compound causal information, and ii) inference from emergent phenomena. Our argument has the following structure:

1. Causal inference from **compound causal information**: We first discuss how causal inference on the aggregate level must be anchored in empirical evidence, and how it can be guaranteed ₅₄₀ that the underlying causal assumptions, modularly encoded in the simulation model, are not distorted in the process of revising the codebase.
2. Causal inference from **emergent phenomena**:
   (a) We go on to explicate an epistemological concept of emergence we find particularly illuminating for our purpose, and we use it to distinguish micro-level and macro-level ₅₄₅ behaviour by drawing on the general scheme developed in Fig. 5.
   (b) We finally bridge the gap between the levels thus introduced again by linking emergent phenomena to the underlying generative rules, thereby once more emphasizing the value of simulation-generated higher-level observations for the understanding of mechanisms in systems biology.

₅₅₀ **Causal inference from compound causal information.** If the goal of modeling a given system is to capture its mechanistic inner workings in a modular fashion, the structured set of $\mathcal{R}$-rules will be as atomic as possible and as basic as needed for the given modeling depth.

Chaining atomic $\mathcal{R}$-rules with causal content will yield "compound" causal information about the underlying causal web, i.e., information about the causal hypothesis or theory (encodable as ₅₅₅ chain or network like in Fig. 3). For this dimension of aggregation, causal inference will be valid if the model is grounded in valid causal knowledge and remains grounded in valid causal knowledge across revisions. We capture this *anchoring requirement* in the twofold condition α:

(α) Empirical knowledge supports both (i) the model's parameters (variables and their values) and (ii) the model's causal structure (i.e., the conditional causal independencies amongst the ₅₆₀ model's variables).

20

The first part of condition α captures the nature of bottom-up modeling, where a model is built on known behaviour of individual agents, i.e. causal knowledge about individual components of the system at the micro-level; the second part is also known as "causal faithfulness" and relates to the macro-level output of the whole system. [24]

To track code revisions, we attach an index to α: α_1 ensures the empirical adequacy of the first simulation model (as a condition attached to the formal scientific model on which the first implementation is based). α_k (for $k = 1, 2, \ldots, k^*$) represents the inductive warrant for valid causal inference along the refinement of the computer program:

(Anchoring) $\mathcal{R}^k$ allows for valid causal inference iff it is *anchored* in empirical knowledge, i.e., iff α_k holds.

This translates directly into conditions for the revision operations: Causal inference on the aggregate level can be called valid across revisions if addition (revision) and deletion (contraction) preserve empirical adequacy both at the micro- and at the macro-level.
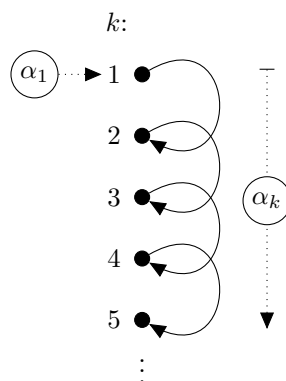

Figure 7: The initial (α_1) and inductive (α_k) warrant for causal inference along the refinement of the computer program $\mathcal{R}^k$ for $k = 1, 2, \ldots, k^*$.

Fig. 7 graphically relates the condition α to the process of revising code-base $\mathcal{R}$. It is important to note that revising $\mathcal{R}$ can be done either (i) by systematically changing the codebase, triggered by mathematical intuition or by accuracy considerations during calibration, or also (ii) by introducing and exploiting new sources of empirical knowledge, triggered by abductive inference or by newly available empirical knowledge. Condition α is meant to track both the empirical adequacy of newly introduced empirical knowledge and the preservation of overall structural integrity (causal

---

[24]Example: Suppose the computer program implements the causal structure $X \to Y \leftarrow Z$ as a set of structural equations of the form $E = f_E(\overrightarrow{C})$ (i.e., the value of the effect variable E is functionally computed from the set of causes $\overrightarrow{C}$). The scientists are now interested in an additional parameter, mediating the influence of X onto Y (the collider node in the corresponding graph). This parameter is added to the underlying assumptions and implemented as U. If causal faithfulness is to be preserved, the modeler has to make sure that both of the following structural assumptions are implemented in $\mathcal{R}$: (i) X still influences Y *indirectly*, and (ii) the variable U screens off Y from X's influence (in accordance with the *Causal Markov Condition*).

faithfulness) even under local modifications.

Causal inference from emergent phenomena. To distinguish the aggregation-based mode of causal inference above from emergence-based inference causal inference, we build on the following two facets of *epistemological emergence* which we find fruitful in the context of this paper:

(E1) *Unpredictability*: Macro-level features of a complex system are emergent if they could not have been predicted from the most thorough knowledge of the system's atoms.

(E2) *Irreducibility*: Macro-level features of a complex system are emergent if they cannot be reduced to fundamental physical theory.

(E1) and (E2) are formulated having in mind physical arrangement or biological (self-)organization and the like (e.g., the formation of crystalline structures). For the purpose of defining what it means to be an emergent feature of a computer simulation, we adjust and combine (E1) and (E2).

A phenomenon observed in a computer simulation is only due to predictable and repeatable calculations, i.e., to simple inference from the computational model (see also Beisbart and Norton, [11], who make this point precise even for Monte-Carlo simulations). Nevertheless, even if computer simulations cannot be understood as unpredictable in this strict sense, due to their narrowly confined sandbox nature, in many instances predicting a simulation's outcome is *practically* impossible due to the complexity of the modeled system (E2). For sufficiently complex systems, predicting the outcome is only possible by stepwise calculation of the system's evolution, which essentially amounts to running the simulation.

Predictability is often contrasted with *surprise*. As mentioned above, Parke ([3]) distinguishes different degrees of surprise a researcher might show when encountering an unexpected result: In some cases, the researcher might even be prompted to revise her initial assumptions. Parke concludes therefore that, even if computer simulations can only be called *simple inference from the computational model*, there is no general epistemic privilege of real experiments over simulations concerning the capacity to generate surprise. What happens in the case study is precisely that the computer simulation points to a gap in the theoretical assumptions underlying the model itself and triggers further refinement. This is particularly evident in the second modelling attempt.

On the basis of these considerations we propose to merge (E1) and (E2), slightly adjusted, into a combined concept of emergence from computer simulation, in line with our case study:

(E*) Macro-level features of computer simulations are emergent i) if they are not describable in lower-level $\mathcal{R}$-terms (i.e., if the higher-level description goes beyond $\mathcal{R}$-vocabulary), ii) if they are not reducible to lower-level $\mathcal{R}$-terms other than by presenting the full protocol of the simulation (or, respectively, an essential part thereof), and iii) if they are not predictable from the components staticly (but only by running the simulation or tracing the protocol).[25]

---

[25]This concept of emergence is related to Bedau's concept of *weak emergence*, where a macrostate of a system

a) MACRO-LEVEL

emergent patterns

generates

MICRO-LEVEL (manipulable through revisions)

s_1 s_2 s_3 s_4 s_5

$\cdots$

evolution of $\mathcal{R}$-parameters
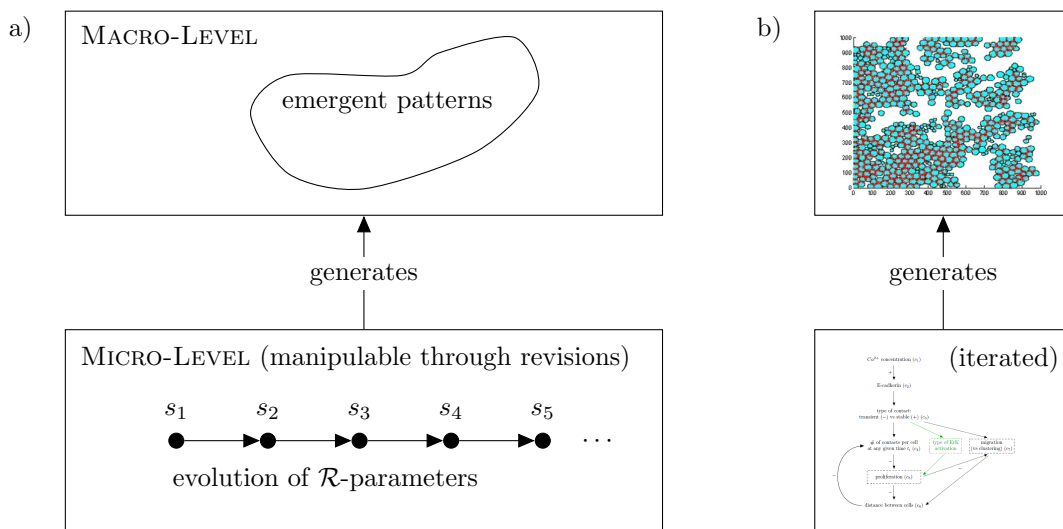
b)

generates

(iterated)

Figure 8: Emergent patterns are induced through the evolution of the implemented system and only indirectly controllable through lower-level parameters. The figure includes material reprinted from [19] under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0).

Fig. 8 illustrates the ingredients and their relations: Fig. 8a abstractly links the evolution of the simulated system $(s_1, s_2, \dots)$ to patterns on a macro-level that are not predictable from the static $\mathcal{R}$-rules or from a snapshot of the $\mathcal{R}$-parameters and not reducible to $\mathcal{R}$-vocabulary, nevertheless generated by aggregation at the lower level. Fig. 8b interprets our case study along the abstract scheme: Iterative calculations of the systems parameters (encoding individual "social behaviour" of the cells) generate structural patterns on the macro-level which are compared at this higher level with empirical data (and found to be in "qualitative agreement" during validation of the third model).

Once condition α inductively warrants valid causal inference from compound causal information across revisions, emergent phenomena (as defined in E*) facilitate causal inference in a special way: not only are emergent phenomena grounded in the atomic rules and their interaction, but they truly only arise in the dynamic evolution of the implemented system.

Since the atomic level is not set up in macro-level terms, though, emergent phenomena cannot be directly subjected to intervention – *in other words, macro-level behavior can only be influenced indirectly by manipulating the basic micro-level.* Yet, although the relation between the codebase

---

is called "weakly emergent" if and only if it can be derived from the microdynamics of this system only by way of simulation ([44]). Our definition goes beyond Bedau's concept, though, in adding language (i.e., level-specific vocabulary) to mark the distinctive gap between what "macro" and "micro" refers to. Explicitly marking this gap allows for distinguishing causal inference as *combining atomic causal assumptions into compound causal claims* from causal inference as *recognizing behavioural patterns on the emergent level.* We believe this is an important distinction since it illuminates how inference from *genuinely* emergent phenomena in our sense may reach beyond language-bound causal learning algorithms.

23

and higher-level observations can be very opaque, emergent phenomena are nevertheless *generated* by aggregation/interaction of atomic rules. So, when running a simulation leads to the detection of emergent events, objects, or patterns, the simulation acts as a *guide to causal understanding*, because – although systematic control over emergent cause-effect relations might not be feasible – perceived macro-level phenomena will possibly suggest discriminating tests (or revisions, respectively) at the atomic level. In this way, emergent phenomena help in tracing the interplay of the target system's mechanisms (and sub-mechanisms): micro-macro interaction.

5. Computer Simulation, Causal Discovery Algorithms, and RCTs

In the following, we want to look at the details of the epistemic role of computer models and simulations in causal inference by comparison with causal learning algorithms and randomized controlled trials. In this comparison, we understand computer simulations along the lines presented above, i.e., simulations based on scientific models of underlying mechanistic relations.

By "causal learning algorithms" we especially refer to automated inference techniques in the spirit of Pearl's IC* algorithm (cf. [1]) or the SGS, PC, and PC* algorithms developed by Spirtes, Glymour, and Scheines (cf. [2]). These procedures were created for the task of 'discovering' causal (Bayesian) networks from a joint probability distribution over a set of variables (joint frequencies, respectively). Their input can be a database, their output will be a network structure, or – to be precise – a class of (observationally equivalent) networks, that can only be narrowed down further if additional knowledge is available.

The third column in the table below looks at randomized controlled trials (RCTs) which have a certain methodological overlap with causal learning algorithms in drawing inferences from statistical data. To clarify this overlap and to make the distinctions between simulation, causal discovery, and RCTs transparent, we investigate the following seven dimensions: (i) the question asked by each method, (ii) the target (output) knowledge, (iii) the required input (and especially the causal information required), (iv) the investigative level(s) at which the method operates, (v) the type of dynamics exemplified by each method, (vi) the methodological characterization distinguishing each method, and (vii) the challenges a researcher is facing in the application of simulations, causal discovery algorithms, and RCTs. Table 1 summarizes the comparison.

Question asked. The three methods can be compared and contrasted by looking at the type of question they answer (the teleological dimension): RCTs ask a very specific question about the relation between two targeted variables, namely about the impact of changing the value of one variable on the value of the second variable (the effect size in the study population, ideally transferable to the target population). Causal discovery algorithms take a broader view and look at the relational structure that can be read off a set of statistical data. And computer simulations really ask (i) about the underlying productive mechanism that supports an observable causal relationship between two (or more) variables, and (ii) the causal dynamics generating the observed phenomenon – in asking "how does X cause Y?", this third type of question requires and aims at knowledge of different quality, as we will discuss next.

Target knowledge. When Cartwright criticizes the formal approaches to causation put forward by Pearl or Spirtes, Glymour, and Scheines, she emphasizes the expressiveness of discipline-specific, connotationally rich terminology used to describe and analyze experiments close to practice. Her

	computer simulation	causal learning algorithms	randomized controlled trials
question asked	what is the underlying mechanism, what are the generative causal dynamics? (micro-macro interaction)	what is the relational structure?	what is the effect (size)?
target knowledge	thick: possible mechanism and dynamics	thin: (class of) network(s)	thin: statistical value
causal input	fundamental mechanistic assumptions	variable selection, distinctness assumption	variable selection, no confounding
inference level	bottom-up (from unit to population)	population level	population level
inference type	dynamic	static	comparative
methodology	generative procedure	global structure analysis	local differential estimate
challenges	causal assumptions justified?	sampling sufficient for analysis?	sampling sufficient for extrapolation?

Table 1: Features of computer simulation, causal learning algorithms, and randomized controlled trials in comparison.

main point is that much of such "thick", useful detail knowledge is lost when formalizing causal relations in purely graph-theoretical, "thin" terms (cf., e.g., [45]). Quite in Cartwright's sense, computational modeling aims at the generation of thick causal knowledge about *how* the parts of a mechanism underlying some phenomenon *possibly* interact to produce the observed behavior *in a dynamic way.* This is in stark contrast to causal discovery algorithms, which are concerned with *actual* data but merely order a set of variables w.r.t. their causal dependencies in thin fashion. The output of an RCT consists in mere difference-making information, i.e., in information about whether and to what degree one variable influences a second. Although this output is only a thin statistical value, the hope is that it might be closer to the ground truth than the information gleaned from discovery algorithms: The RCT was carefully designed to isolate the targeted relation, whereas we might not have much information about the quality of the database the discovery algorithm processes (see also [46] for the related discussion about the distinction between real and virtual clinical trials).

Causal input. All three methods can be used for causal inference in one way or another, but interestingly, all three also require causal pre-knowledge as input in one way or another: Computer simulations are based on fundamental mechanistic assumptions and allow for causal insights at the macro-level, based on causal knowledge of the micro-level (see our anchoring conditions above). A similar finding is true for causal discovery algorithms: The input, a frequency or probability table, encodes basic causal assumptions in providing information about variables assumed to be causally sufficient for describing the targeted phenomenon and sufficiently distinct to permit representation in an interventionist Bayes net framework (with the Markov condition being a central modeling

25

criterion). In contrast, causal pre-knowledge plays only a minor role in designing and evaluating RCTs: The goal of RCT methodology is precisely to avoid causal assumptions as much as possible and neutralize (or minimize), through randomization, all potential confounding factors. Nevertheless, implicit causal pre-knowledge enters in the choice of the variables and in their definition (e.g., the causal fact that all confounders are included in the variable set), and more obviously, in extrapolation.

Inference level. In this respect, computer simulation and RCTs stand in clear contrast: If the computational model encodes fundamental mechanistic entities and relations, the computer simulation takes the atom (or the unit) as a starting point and arrives at the population level in a bottom-up fashion, whereas RCTs necessarily operate at the population level, because they essentially require randomization and averaging. Causal discovery algorithms are agnostic as to at what level the input data are harvested, but in the context of systems biology or pharmacological research, the data typically capture population-level probabilities (i.e., frequencies) only.

Inference type. As we showed in our case study above, computer simulation techniques can truly be called *dynamic* inference methods: They simulate the dynamic behavior of a population by iteratively updating the virtual version of the targeted system (for an arbitrary amount of steps). Inference would not be possible without this dynamic property. They can also be called *dynamic* in an additional respect: For the purpose of validation, the computational model must be revised dynamically in order to increase the adequacy of the model and the predictive accuracy of the simulation. Causal discovery algorithms, on the other hand, can be called *static* in all respects: They take as input a static table and deductively compute a structural representation of this input. Our third causal inference method, RCTs, might be called *comparative*: The effect size as output is evidently calculated from statistical data, but much effort goes into defining the control and the treatment group (to be understood as an *intervention* in the population) in order to bring out the targeted contrast.

Methodology. As we illustrated in the points above, computational simulation is indeed a *generative procedure*: It generates dynamic behavior from fundamental units and thereby generates a possible picture of the mechanism responsible for the observed phenomenon. In doing so, the simulation provides a *how-possibly* explanation whose strength must be tested in validation cycles. Causal discovery algorithms produce structures, but we would not call these algorithms "generative procedures", since they deductively calculate a network from a database. In this sense, they should rather be called "structure analysis", since the graphical structure they return as an output is contained in the probabilistic dependencies and independencies between the variables in the database. In contrast with RCTs, such causal discovery algorithms look at a set of variables in a 'global' way. RCTs focus on a specific causal link in a 'local' way – 'local' because (i) they are concerned with the (differential) influence of one variable on a second, and (ii) because they are conducted for a local set of population characteristics.

Challenges. All three inference methods come with their own challenges that are closely related to the question of how their results can be validated. Computer simulations face the problem of justifying the causal assumptions on which their simulated evolutions rest. We showed above how the computational model as a theoretical basis can be validated in the cycle of simulation, experiment, and revision. In the case of causal discovery algorithms, the question arises whether the statistical data are sufficient for an analysis (in narrowing down the output class of networks

26

as far as possible), but validation as such is not in the focus, because such discovery algorithms
arrive at their results in a deductive way. One can of course ask additional questions about the
data source, e.g., if the sampling is a valid representation of a larger population. This is precisely
the kind of question, an RCT must answer before its results can be extrapolated: Once it is proven
internally valid (i.e., the causal relation is indeed sufficiently isolated), the external validity of the
trial must be checked by evaluating its representativeness (of a target population).

This comparative list makes the distinctive epistemic role of modeling and simulation transparent and conceptually demarcates its inferential process from other techniques aimed at learning
about the causal features of our world. The added value of computational simulation consists in
going beyond the phenomenological surface and providing insights into lower-level mechanistic processes. As a generative procedure, this technique certainly demands quite some scientific creativity
of the researcher, but this precisely unlocks the capacity for scientific discovery from emergent
phenomena. As a bottom-up constructive method, computational simulation is able to deal with
complex phenomena which otherwise evade straightforward data-driven analysis. Nevertheless, this
complexity and the intricate relation between unit and population-level behavior need not remain
obscure, as we argued above by formulating our refined understanding of emergence.

6. Causal Inference from Modeling and Simulation

The philosophical literature on modelling is increasingly vast, however clear formal analyses
of computational modelling in systems biology are still lacking. In many instances, the inferential
power of computer simulations for causal assessment is called into question by comparing simulations
to real experiments – in a misleading way, we think. As quoted above, Diez Roux summarizes her
discussion of the epistemic value of computer simulations as follows ([43, p. 101]):

> In the simulated world, everything is counterfactual in the sense that the world and all
> possible scenarios are artificially created by the scientist.

As a possible response to the charge of vacuousness, natural variability might be mimicked by
introducing randomness in the computational model, as, e.g., in the case of Monte Carlo simulations.
Yet again, are those simulations capable of delivering truly *novel* results? Beisbart and Norton
([11]), for example, dismiss the idea of computer simulations as representing any kind of experiment.
Their claim: Computer simulations represent inferences and not experiments, namely inferences
"by ordinary argumentation of conclusions from assumptions built into the simulations" (cf. [11,
p. 403]). In their view, this extends to Monte Carlo simulations:

> It is true that Monte Carlo simulations appear, initially, to be quite different from other
> simulations because of the essential role of randomness. This appearance is deceptive,
> for randomness can have no role or only a minimal one in the final results if they are to
> be credible. Otherwise Monte Carlo simulations would be little better than forecasting
> by reading tea leaves or shuffling tarot cards, whose results are mostly expressions of
> random chance. The results of a good Monte Carlo simulation are not. They are
> repeatable and reliable.

And they go on to comment on the idea of gathering novel evidence from computer simulations:

27

In real experiments, the unpredictable and novel character of the data is the source of
⁷⁸⁰ the final experimental result. In a Monte Carlo simulation, the unpredictable and novel
character of the random numbers is minimized and plays almost no role in the final
result.

In other words, a phenomenon observed in the computer simulation is nothing but simple inference
from the computational model. Nevertheless, to draw the conclusion from these observations that
⁷⁸⁵ computational modeling cannot contribute to causal learning, would be false, as we argue with this
paper: We emphasize the iterated interplay between scientific knowledge as input, modeling and
simulation, and revision of the computational model in the dialog between virtual and experimental
data-gathering, geared towards validation of the scientific assumptions. In our comparison of mod-
eling and simulation with causal discovery algorithms and RCT methodology (Table 1), we carved
⁷⁹⁰ out the distinctive epistemic features and specific advantages of agent-based computer simulations:
The bonus is twofold – not offers modeling and simulation insights into the phenomenological black-
box, it also allows tracing the causal dynamics at work.

As an illustration of these epistemic advantages, we took as a case study, the discovery of the
⁷⁹⁵ functional properties of a specific protein, E-cadherin, which seems to have a key role in metastatic
processes by way of influencing cell growth and proliferation signalling. The case study was pre-
sented in three stages, following the development of the computational model from its purely agent-
based bottom-up version to a refined, multi-scale implementation and on to a final single-scale,
agent-based formulation with the local insertion of a phenomenologically motivated rule (repre-
⁸⁰⁰ senting the exclusive role of temporal contact length in reproducing the growth patterns observed
experimentally).

To capture the model's development, we followed Margaret Morrison in emphasizing the role of
the computational model as object of inquiry, to be subjected to formal experimentation. The case
⁸⁰⁵ study is particularly interesting in this respect, because reconstructing the process revealed both
the chosen strategy towards validating the computational model and the value of the intrinsic link
between the empirically grounded modular codebase and observations at the aggregate level.

A central part of our general reconstruction of a computer simulation's design process is the
⁸¹⁰ simulation–validation–revision cycle. In order to fruitfully integrate the phases of this cycle, we
supplemented our scheme with formal notation as a potential basis of a protocol for valid scien-
tific inference from computer simulations. By defining an anchoring condition, we explicated the
formation of empirically grounded aggregate knowledge from atomic information. We sketched a
robustness measure for evaluating a simulation's quality in terms of multi-dimensional accuracy and
⁸¹⁵ utilized belief revision terminology to specify how the codebase may be changed while preserving
empirical adequacy (via anchoring).

We ultimately arrived at an analytic characterization of emergence that goes beyond outlining
semantic constraints for scientific talk about emergent phenomena – our characterization demarcates
⁸²⁰ different information-bearing levels and bridges this gap in showing how observations of emergent
phenomena can fruitfully be utilized for directly testable causal inference despite their higher-
level mediate nature. In lining up all ingredients for a validation–revision protocol we illuminated

how well-designed computer simulations – though lacking the material link to their target systems as virtual, artificial constructs – are justified guides to a mechanistic understanding of biological systems.

Appendices

A. Rules Dictating Cell Behaviour

The set of relevant actions for cells in the model are: 1) bonding (to other cells and to substrate), 2) spreading (on the substrate), 3) migration (once bonded to the substrate), 4) apoptosis (if failure to bond to substrate takes place), 5) cell growth, and 6) division (mitosis). In addition to the internal parameters, each cell is assigned a set of parameters that defines the exact location of its centre in x, y, and z Cartesian coordinates. The agents (the cells) progress through the cell cycle in a sequence of phases: the interphase (I) – which comprises the growth phase 1 (G1), the synthesis phase (S), and the growth phase 2 (G2) – and the division (mitosis, M) and reproduction phase; the G0 phase represents a state of "quiescence" where the cell is neither dividing itself, nor preparing to divide, but merely performing its biological functions (see Fig. A.9).
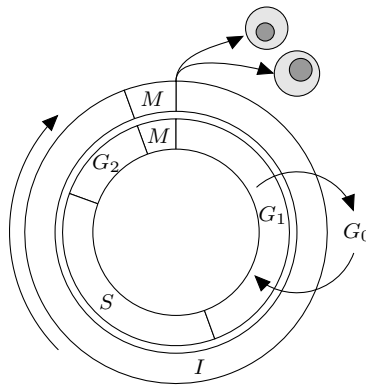

Figure A.9: Cell cycle: cell growth and division consist of two main phases: the interphase (I) during which the cell grows and synthesises a copy of its DNA in order to replicate, and the mitotic phase (M) during which the cell divides up into two daughter cells. The interphase is subdivided into the G1 phase, that is the first growth phase; the synthesis phase (S) during which its DNA is replicated in the nucleus, and the G2 phase during which it further grows before it enters the mitotic phase.

The ABM rules governing cell progression and proliferation are formalized as illustrated in Tab. A.2 (cf. [17]) and related to biological behaviour below.

Bonding. The rule for bonding is based on published data ([47]), according to which the probability of two adjacent cells bonding to each other is inversely proportional to their distance, and directly proportional to the concentration of calcium ions in the culture (see fig. 2).

CURRENT STATE		ACTION
Cell is in G1		Increase cell volume by set amount; Increment internal counter by +1
Cell is at G1/G0 check-point	if cell is bonded to 4 or more neighbours	Enter G0
	If cell is not sufficiently spread (radius <1.5x 'rounded' radius)	
	ELSE	Continue in G1
Cell is in G0	If cell is bonded to less than 4 neighbours	Re-enter G1
	If cell is sufficiently spread	
	ELSE	Remain in G0
Cell is in S/G2		Increment internal counter by +1
Cell is at G2/M boundary		Change to mitotic cell type
Cell is at end of M phase		Divide to create two identical daughter cells

Table A.2: Summary of the Agent Based Model rules where agents represent cells and their actions in the cell cycle phases.
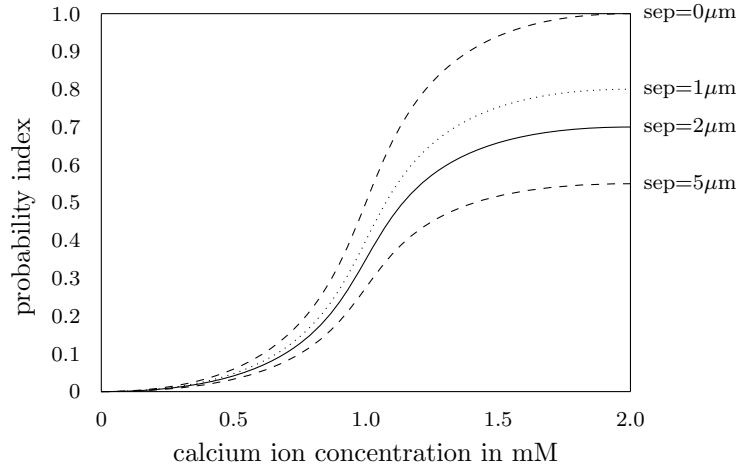

Figure A.10: Probability densities corresponding to four cell:cell distance ranges $[0\mu m, 1\mu m, 2\mu m, 5\mu m]$ as a function of the Calcium concentration in mM (cf. [18])

Spreading. Cells need to bond to the substrate in order to proceed in cycle progression, otherwise they enter into apoptosis; in order to bond to the substrate, they form numerous attachments thus assuming a flattened morphology. "This process is *emulated* in the model by increasing the

surface area of attached cells by a small amount at each iteration, whilst maintaining a constant volume. Cells continue to spread until their radius has increased by a factor of approximately 1.5" ([17], see also [35]).

Migration. Migration is also modelled on the basis of experimental observations: the allowed migration distance, 2.5 times the cell radius, was obtained from time-lapse recording of a wound closure assay. Direction of movement is chosen randomly within a 60° angle of the original trajectory; and if movement in the chosen direction is blocked by another cell, then the direction is changed up to 10 times, after which the cell will remain stationary.

Apoptosis. Cells that fail to bond to the substrate before the inception of the first growth phase (G1) die (they enter apoptosis). Dead cells are known to be fragmented and phagocytosized by their neighbours. In models of cell growth, dead cells shrink and are removed from the model. However, for specific epithelial cells, such as urothelial cells, apoptosis rarely applies, and therefore this rule is not implemented in models for this kind of cells, as the one considered here.

Mitosis. The final cell cycle phase results in cell division and replication. During cell division (mitosis), cells have been observed to round up and thus lose adhesion with their neighbours and with the underlying substrate to some extent; hence model cells undergoing the mitotic phase are assumed to alter their morphology from ellipsoidal to spherical and to break up all their bonds with neighbouring cells, eventually dividing up into two spheres. New cells are spherical and have no bonds with other cells. Furthermore, they do not form new bonds or migrate until mitosis is complete.

From the point of view of modeling cell proliferation, "quiescence" and "contact inhibition" are also essential.

Quiescence. The cell cycle has a checkpoint, a sort of "bifurcation point" after which, based on its morphology at that given time, and on the number of contacts the cell has formed with neighbouring cells, it either proceeds through the remainder of the cell cycle and reproduces, or becomes quiescent, that is, stops proliferating. Quiescence may be caused by "contact inhibition"; this means that, after a certain amount of contacts with other cells, a cell stops proliferating.

Contact Inhibition. In the model, the contact inhibition rule dictates that a cell becomes quiescent once it has ≥ 4 bonds with neighbouring cells (or, alternatively, if it has 3 bonds and is adjacent to the "edge" of the substrate). Quiescent cells remain in this state until enough of their intercellular bonds are broken so as to allow them to re-enter the cycle. This rule also follows observations showing that in the culture of epithelial cells, those situated at the centre of the colonies *in physiological or high calcium media* withdraw from the cell cycle and become quiescent. Since physiological calcium is necessary for calcium-dependent physical bonds such as those created by E-cadherin, it has been hypothesized that cells become "aware" of the presence of nearest neighbours through intercellular signalling, via E-cadherin bonds forming the intercellular matrix.

At each tick of the simulation, each cell is interrogated and its response depends on the combination of states of all the above parameters.[26] The formation of extracellular E-cadherin bonds depends on the presence of calcium ions, therefore, the action of intercellular signalling on cellular growth is studied in experiments by putting cells in cultures with a physiological (around 2mM) vs. a low concentration of calcium ions (around 0.09mM), which is assumed to create a situation of higher vs. lower number of intercellular contacts (and, therefore, of growth inhibition vs. proliferation, respectively).

The simulation was run under the two conditions of high vs. low calcium environment (with related bonding probability) in order to verify whether these different settings would reproduce the kind of distinct behavioural patterns observed experimentally in high vs. low calcium environments.

B. The Causal Structure Underpinning the Set of Rules

The assumed causal structure underlying the dynamics of cell growth through "contact inhibition" is presented in the left diagram of Fig. 3. The graph represents the hypothesized causal structure underpinning the phenomenon of cell proliferation as higher-level behaviour resulting from cell:cell interactions; this can be considered as the theory that needs confirmation through modelling, simulation and comparison with experimental data ("validation"). The main causal variables in the structure are:

c_1) the characteristics of the environment in terms of Calcium ions (Ca^{2+}) concentration (physiological vs. low calcium condition);

c_2) quality and functionality of the E-cadherin present in the cell culture as a consequence of the concentration of calcium ions;

c_3) type of cell:cell contacts: stable (as promoted by the presence of E-cadherin) vs. transient;

c_4) number of cell:cell contacts (for any given cell at any given time the cell goes into quiescence and ceases to proliferate whenever it is in contact with ≥ 4 other cells);

c_5) cell proliferation as a result of the division and replication of each cell in the population;

c_6) distance between cells as a result of the topology determined by cell proliferation; reduced distance increases the probability of contacts which in turn determine quiescence (i.e., inhibit proliferation) when they reach the number of four in each cell;

c_7) migration is inhibited both by cell:cell contacts and is made more and more difficult as the population grows; in turn, migration also contributes to determining the distance between cells by changing the topology of the cell population at each iteration.

The culture medium (high vs. low concentrations of Ca^{2+} ions) determines a higher vs. low capability of E-cadherin to form contacts, which in turn determines the presence of intercellular

---

[26]A numerical technique is applied to correct for any overlap between neighbouring cells that may have resulted from cell growth, spreading, or division during the previous agent based iteration: these may consist for instance in identifying any overlap between every pair of neighbouring cells and applying a force proportional to this overlap to separate them ([16]).

junctions, contributing to stable vs. transient contacts respectively. The *number* of contacts that any given cell has at any time is obviously a function of how stable they have been up to then. In turn, a high number of contacts (≥ 4) signals the cell to stop proliferating.

⁹²⁰ Proliferation is part of a negative feedback in that the higher the proliferation, the higher the population density, and therefore the shorter the distance among cells on average, which leads to a higher probability of cell:cell contacts and therefore to inhibition of proliferation. Proliferation also negatively impacts on migration (high density of cells inhibit their movement), whereas the presence of contacts inhibits migration (and thereby promotes clustering).
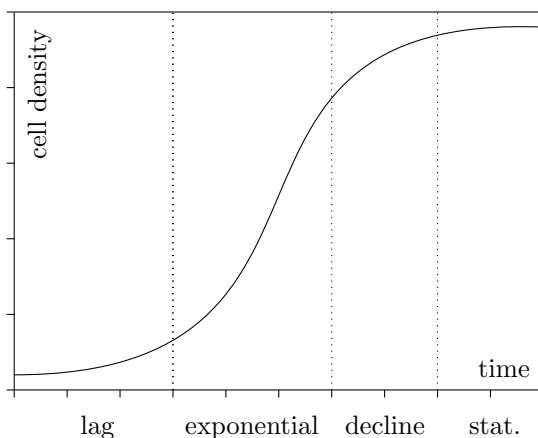
⁹²⁵ **C. Cell Growth Benchmark**

Figure C.11: Cell growth function with lag phase, exponential phase, decline, and stationary phase.

D. Modifications of the ABM component for the Second Model

In the second model, the previous ABM component is not only integrated with an ODE system, but is also modified as follows:

E-cadherin mediated contacts. Whereas in the previous model, bonding probability was ⁹³⁰ straightforwardly determined by the concentration of calcium ions in the medium and by the distance among cells, in the new version of the model the rule is expanded so as to qualitatively reflect more adequately observations in [47, 48]. In addition to calcium concentration in the medium and distance between the cells, bonding probability is also determined by "adhesion probability", which is a function of the cell E-cadherin expression level: this varies between 1 (normal expression) and ⁹³⁵ 0 (no expression, e.g., in malignant cells).

33

Transient cell:cell contacts. Whereas the first model only took into account the number of contacts as a basis for determining cell cycle progression, the new one also accounts for the contact *mode* (stable vs. transient), and allows for the formation of prolonged contacts also in low calcium concentration media (although significantly less in number). This makes it possible for cell:cell signalling to occur also through non-E-cadherin mediated contacts, although much less frequently so.[27] Moreover, E-cadherine mediated contacts may be also formed in low calcium concentrations environments, but, according to the bonding probability rule (see previous paragraph and fig. 3), they too are much less frequent than in environments with physiological concentrations of calcium.

Migration rules. Whereas in the first model, cells where allowed to migrate only when they had 0 contacts with other cells, in the second one, migration results from the difference of strength between the "active migration force" possessed by a cell, and the intercellular "adhesion force" which binds it to other cells (this being a function of the stability/transiency of the contact, itself mediated by E-cadherin expression levels). When the net adhesion force is less than 20nN, then the cell migrates.

E. Testing the Modelling Assumptions of the Second Model

When the behavior displayed by a simulation model contradicts the expected results, then various, *non mutually exclusive* sources of discrepancy can be taken to account for this mismatch (see Sec. 4.3 for a systematic discussion of such sources of discrepancy). Generally, a broad distinction between "internal validity" (generally referred to as "model verification") and external validity ("model validation") is drawn. The former regards the formal (mathematical/logical) consistency of the model itself; the latter its empirical adequacy. Scientists generally proceed by first testing the former, and, if no sources of error are found, then they go on to challenge the theoretical model underpinning the computational one, and change it more or less drastically, on the basis of explanatory hypotheses for the obtained results, and of new empirical knowledge possibly acquired in the meanwhile.

In this "Lakatosian" process, changes to the model follow an ordering which depends on the steadiness of the theoretical/empirical knowledge on which the model is based. The parts of the model which are more insecure are tested for their individual theoretical solidity and/or empirical adequacy; if these "pass the test", then other parts of the model are put to the test. However, given the holistic aspect of computational modeling and simulation, it may happen that something goes wrong with the overall simulation, even if all its components "pass the test"; in this case the problem may lie in how such components are combined – this is especially true for multi-scale models, because of possible inconsistencies arising in the transformation of one submodel's outputs into the other submodel's inputs.

Instead, if the revised model delivers the expected results, then this is considered to provide evidence that the mechanisms underlying the target system it represents are analogous to the rules

---

[27]Furthermore, whereas the old model only took into account the number of "non-dimensional" contacts, the new model also takes into account the physical length of the bond. The rationale for this rule would be that the longer the bond, the more receptors and ligands enter into contact and activate the intracellular signalling pathway. This aspect of cell:cell signalling is taken into account in Eq. 6 of the ODE model. However, as the individual cell:cell simulation will show, only the temporal length of the contact is relevant for patterns of Erk activation, whereas the spatial length is irrelevant. See Sec. 3.2.

underpinning the simulation model. How exactly?

The final part of the case study will help us identify the essential epistemic tools used in model revision and confirmation. These will be formally explicated in the third section of the paper.

Test 1: Sensitivity analysis of the mathematical model. In our case, the scientists started by doing a sensitivity analysis of the mathematical model representing the signaling pathway:[28] each of the 50 parameters in the model was systematically varied by first dividing, then multiplying the value of each parameter by a factor of two.[29]

Although both amplitude and duration of the Erk-PP signals varied when the kinetic parameters (rate constants) were altered, none of the parameter variations affected the fundamental result that "transient" contacts produce a relatively large but transient response, compared to the smaller but sustained response associated with slowly-growing stable contacts. Hence, the simulation result could not be ascribed to the particular contingent combination of parameter values that underpinned it in that specific setting. Interestingly, the *duration of the signals* were relatively insensitive to the model parameters: this means that the *temporal length of the signal, rather than its intensity*[30] was a robust predictor of the model outputs.
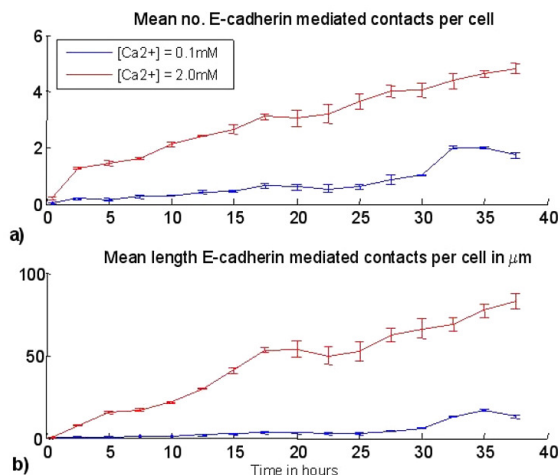

Figure E.12: Mean number (a) and mean total perimeter length (b) of E-cadherin mediated contacts per cell in high calcium (red) vs. low calcium (blue) environment. Figure reprinted from [19] under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0).

---

[28]The signaling pathway model was solved both for a "transient" contact (with a fixed size of 15μm) and for a "stable" contact (starting with 3μm size and constantly growing).

[29]The rationale for testing the model for such a large perturbation was theoretically/empirically grounded on previous studies related to rate constants for binding, dissociation and internalization of epidermal growth factors ([39]).

[30]I.e., number of involved receptors and ligands on the cell surface at any point in time.

35

Test 2: Average number and spatial length of E-cadherin mediated contacts in ABM. Another verification test regarded the information provided by the ABM component of the model to the mathematical model: number and length of the intercellular contacts in the two environmental conditions. This served the purpose of establishing whether through iterations, the ABM component would systematically deliver different initial conditions to the mathematical model depending on whether the simulation was run in low vs. high calcium conditions. The test showed indeed different patterns of growth in the number and in the perimeter of E-cadherin mediated contacts in the two settings (see Fig. E.12).

Also tests for empirical adequacy were carried out. These served the purpose of verifying the empirical adequacy of at least some parts of the model.

Test 3: partial validity test for proportion of E-cadherin mediated contacts in the two settings. The revised version of the ABM component in the model prescribed "more realistic rules" for the formation of E-cadherin bonds, e.g., by allowing these to be formed also in low calcium conditions (although less frequently so). The incidence of E-cadherin mediated contacts in the two simulated settings was found to be in good qualitative agreement with the distribution of cells with low/moderate/high Erk activation in experimental settings. This is not a direct validity test, but one which draws on explanatory reasoning: The immunofluorescence of the distribution of Erk-PP in urothelial cells maintained in low vs. physiological Ca^{2+} conditions shows that in all cells there is some degree of Erk activation, but that there is also a clear heterogeneity between levels of Erk activation in the two cultures. Hence it is suggested that, in addition to the EGF stimulation explaining the activation of Erk tout court, the contact *mode* might be invoked as a possible explanation for such an heterogeneity of activation levels. And the most proximate cause for such qualitative difference is traced back to stable vs. transient contacts, as determined by the presence vs. absence of E-cadherin respectively.

Test 4: validity test for the topological distribution of E-cadherin mediated contacts in the two settings. Fig. E.13 show the qualitative agreement between the distribution of E-cadherin after 50 iterations in the simulated high vs. low calcium settings. These agree qualitatively with the microscopy images of normal urothelial cell cultures cultured in low vs. physiological calcium environments.

Test 1 establishes that the ODE system is working properly, and test 2 says that the ABM component is delivering the right inputs. So maybe the problem lies in the combination of the two. Tests 3 and 4 show that the role of low vs. high physiological calcium levels for the formation of E-cadherin and for Erk activation agree in the source and the target system.

These constraints, together with the uncontested role of contact inhibition and with a deeper analysis of the inconsistency between the individual level and the population level simulation, provide the basis for the third model.

References

[1] J. Pearl, Causality: Models, Reasoning, and Inference, 1st Edition, Cambridge University Press, 2000.
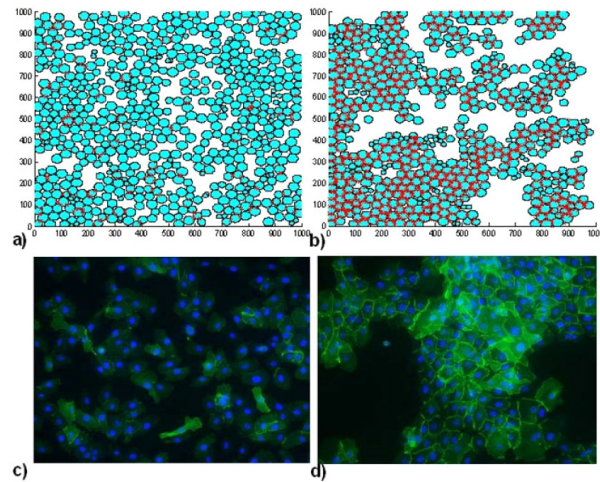
Figure E.13: Comparison of the distribution of E-cadherin in low vs. high calcium environment in computer simulation (a vs. b) and microscopy images (c vs. d). In the simulation, E-cadherin mediated contacts are in red; in the microscopy images, they are identified by immunofluorescence: cells are labelled with an antibody that specifically recognises E-cadherin (green areas of the picture). Figure reprinted from [19] under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0).

[2] P. Spirtes, C. Glymour, R. Scheines, Causation, Prediction, and Search, Adaptive Computation and Machine Learning, 2000.

[3] E. C. Parke, Experiments, simulations, and epistemic privilege, Philosophy of Science 81 (4) (2014) 516–536.

[4] W. S. Parker, Does matter really matter? computer simulations, experiments, and materiality, Synthese 169 (3) (2009) 483–496.

[5] L. Casini, G. Manzo, Agent-based models and causality – a methodological appraisal (unpublished manuscript).

[6] E. Winsberg, A tale of two methods, Synthese 169 (3) (2009) 575–592.

[7] M. Weisberg, Simulation and Similarity: Using Models to Understand the World, Oxford University Press, 2013.

[8] R. Dardashti, K. P. Y. Thébault, E. Winsberg, Confirmation via analogue simulation: What dumb holes could tell us about gravity, British Journal for the Philosophy of Science.

[9] R. Dardashti, S. Hartmann, K. P. Y. Thébault, E. Winsberg, Confirmation via analogue simulation: A bayesian analysis (2015).

[10] R. Frigg, J. Reiss, The philosophy of simulation: hot new issues or same old stew?, Synthese 169 (3) (2009) 593–613.

[11] C. Beisbart, J. D. Norton, Why monte carlo simulations are inferences and not experiments, International Studies in the Philosophy of Science 26 (4) (2012) 403–422.

[12] M. Morrison, Reconstructing Reality: Models, Mathematics, and Simulations, OUP, 2015.

[13] P. Humphreys, The philosophical novelty of computer simulation methods, Synthese 169 (3) (2009) 615–626.

[14] M. Viceconti, A tentative taxonomy for predictive models in relation to their falsifiability, Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 369 (1954) (2011) 4149–4161.

[15] J. O. Dada, P. Mendes, Multi-scale modelling and simulation in systems biology, Integrative Biology 3 (2) (2011) 86–96.

[16] D. C. Walker, J. Southgate, The virtual cella candidate co-ordinator for middle-outmodelling of biological systems, Briefings in bioinformatics (2009) bbp010.

[17] D. C. Walker, J. Southgate, G. Hill, M. Holcombe, D. R. Hose, S. M. Wood, S. Mac Neil, R. H. Smallwood, The epitheliome: agent-based modelling of the social behaviour of cells, Biosystems 76 (1) (2004) 89–100.

[18] D. C. Walker, G. Hill, S. M. Wood, R. H. Smallwood, J. Southgate, Agent-based computational modeling of wounded epithelial cell monolayers, NanoBioscience, IEEE Transactions on 3 (3) (2004) 153–163.

[19] D. C. Walker, N. T. Georgopoulos, J. Southgate, From pathway to population – a multiscale model of juxtacrine egfr-mapk signalling, BMC Systems Biology 2 (1) (2008) 1–17.

[20] N. T. Georgopoulos, L. A. Kirkwood, D. C. Walker, J. Southgate, Differential regulation of growth-promoting signalling pathways by e-cadherin, PLoS One 5 (10) (2010) e13621.

[21] D. C. Walker, N. T. Georgopoulos, J. Southgate, Anti-social cells: predicting the influence of e-cadherin loss on the growth of epithelial cell populations, Journal of theoretical biology 262 (3) (2010) 425–440.

[22] J. Southgate, K. A. Hutton, D. F. Thomas, L. K. Trejdosiewicz, Normal human urothelial cells in vitro: proliferation and induction of stratification, Laboratory Investigation 71 (4) (1994) 583–94.

[23] R. Frigg, Fiction in Science, in: Fictions and Models: New Essays, Munich: Philosophia Verlag, 2010.

[24] R. Frigg, Models and fiction, Synthese 172 (2) (2010) 251–268.

[25] A. Toon, The ontology of theoretical modelling: Models as make-believe, Synthese 172 (2) (2010) 301–315.

[26] P. Teller, Fictions, Fictionalization, and Truth in Science, in: Fictions in Science. Philosophical Essays on Modelling and Idealisation, London: Routledge, 2009.

[27] N. Cartwright, The Dappled World: A Study of the Boundaries of Science, Cambridge University Press, 1999.

[28] S. M. Downes, The importance of models in theorizing: A deflationary semantic view, PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1992 (1992) 142–153.

[29] R. Frigg, Scientific representation and the semantic view of theories, Theoria 21 (1) (2006) 49–65.

[30] J. D. Norton, Thought Experiments in Einstein's Work, in: G. J. M. T. Horowitz (Ed.), Thought Experiments In Science and Philosophy, Rowman and Littlefield, 1991.

[31] J. Reiss, Causal inference in the abstract or seven myths about thought experiments, in: Causality: Metaphysics and Methods Research Project, LSE, 2003.

[32] R. Sorensen, Thought Experiments, New York: Oxford University Press, 1992.

[33] S. Hartmann, The world as a process: Simulations in the natural and social sciences, in: R. Hegselmann (Ed.), Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View, 1996.

[34] P. Humphreys, Extending Ourselves: Computational Science, Empiricism, and Scientific Method, Oxford University Press, 2004.

[35] C. M. Nelson, C. S. Chen, Cell-cell signaling by direct contact increases cell proliferation via a pi3k-dependent signal, Febs Letters 514 (2) (2002) 238–242.

[36] F. A. Brightman, D. A. Fell, Differential feedback regulation of the mapk cascade underlies the quantitative differences in egf and ngf signalling in pc12 cells, FEBS letters 482 (3) (2000) 169–174.

[37] B. N. Kholodenko, O. V. Demin, G. Moehren, J. B. Hoek, Quantification of short term signaling by the epidermal growth factor receptor, Journal of Biological Chemistry 274 (42) (1999) 30169–30181.

[38] C. L. Adams, Y.-T. Chen, S. J. Smith, W. J. Nelson, Mechanisms of epithelial cell–cell adhesion and cell compaction revealed by high-resolution tracking of e-cadherin–green fluorescent protein, The Journal of cell biology 142 (4) (1998) 1105–1119.

[39] C. M. Waters, K. C. Oberg, G. Carpenter, K. A. Overholser, Rate constants for binding, dissociation, and internalization of egf: Effect of receptor occupancy and ligand concentration, Biochemistry 29 (14) (1990) 3563–3569.

[40] A. E. DeWitt, J. Y. Dong, H. S. Wiley, D. A. Lauffenburger, Quantitative analysis of the egf receptor autocrine system reveals cryptic regulation of cell response by ligand capture, Journal of Cell Science 114 (12) (2001) 2301–2313.

[41] D. A. Lauffenburger, G. T. Oehrtman, L. Walker, H. S. Wiley, Real-time quantitative measurement of autocrine ligand binding indicates that autocrine loops are spatially localized, Proceedings of the National Academy of Sciences 95 (26) (1998) 15368–15373.

[42] G. Oehrtman, H. Wiley, D. Lauffenburger, Escape of autocrine ligands into extracellular medium: experimental test of theoretical model predictions, Biotechnology and bioengineering 57 (5) (1998) 571–582.

[43] A. V. Diez Roux, The virtual epidemiologist – promise and peril, American Journal of Epidemiology 181 (2) (2015) 100–102.

[44] M. A. Bedau, Weak emergence, Philosophical Perspectives 11 (s11) (1997) 375–399.

[45] N. Cartwright, Causation: One word, many things, Philosophy of Science 71 (5) (2004) 805–819.

[46] B. Osimani, R. Poellinger, M. Bertolaso, Real and Virtual Clinical Trials: a Formal Analysis-Forthcoming.

[47] W. Baumgartner, P. Hinterdorfer, W. Ness, A. Raab, D. Vestweber, H. Schindler, D. Drenckhahn, Cadherin interaction probed by atomic force microscopy, Proceedings of the National Academy of Sciences 97 (8) (2000) 4005–4010.

[48] Y.-S. Chu, W. A. Thomas, O. Eder, F. Pincet, E. Perez, J. P. Thiery, S. Dufour, Force measurements in e-cadherin–mediated cell doublets reveal rapid adhesion strengthened by actin cytoskeleton remodeling through rac and cdc42, The Journal of cell biology 167 (6) (2004) 1183–1194.