Systems biology

Complete populations of virtual patients for *in silico* clinical trials

S. Sinisi [1], V. Alimguzhin [1], T. Mancini [1]*, E. Tronci [1], B. Leeners [2]

[1]Computer Science Department, Sapienza University of Rome, Italy
[2]Division of Reproductive Endocrinology, University Hospital Zurich, Switzerland

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation. *Model-based* approaches to safety and efficacy assessment of pharmacological drugs, treatment strategies, or medical devices (*In Silico* Clinical Trials, ISCT) aim to *decrease* time and cost for the needed experimentations, *reduce* animal and human testing, and enable *precision medicine*. Unfortunately, in presence of *non-identifiable* models (*e.g.*, reaction networks), parameter estimation is not enough to generate *complete* populations of Virtual Patients (VPs), *i.e.*, populations guaranteed to show the *entire* spectrum of model behaviours (*phenotypes*), thus ensuring *representativeness* of the trial.

Results. We present methods and software based on global search driven by statistical model checking that, starting from a (non-identifiable) quantitative model of the human physiology (plus drugs PK/PD) and suitable biological and medical knowledge elicited from experts, compute a *population of VPs* whose behaviours are *representative* of the whole spectrum of phenotypes entailed by the model (*completeness*) and *pairwise distinguishable* according to user-provided criteria. This enables *full granularity control* on the size of the population to employ in an ISCT, *guaranteeing representativeness while avoiding over-representation of behaviours*.

We proved the effectiveness of our algorithm on a non-identifiable ODE-based model of the female Hypothalamic-Pituitary-Gonadal axis, by generating a population of 4 830 264 VPs stratified into 7 levels (at different granularity of behaviours), and assessed its representativeness against 86 retrospective health records from Pfizer, Hannover Medical School and University Hospital of Lausanne. The datasets are respectively covered by our VPs within Average Normalised Mean Absolute Error of 15%, 20%, and 35% (90% of the latter dataset is covered within 20% error).

1 Background

Model-based approaches to safety and efficacy assessment of drugs, pharmacological treatments, or medical devices (*In Silico* Clinical Trials, ISCT) hold the promise to *decrease* time and cost for the needed experimentations, *reduce* the need for animal and human testing, and enable *precision medicine*, where personalised treatments or devices *optimised* for each patient can be designed before being actually administered or implanted (Avicenna Project, 2016; Pappalardo *et al.*, 2019). To enable ISCT, *quantitative mechanistic models* (Virtual Physiological Human, VPH, models) of the human (patho-) physiology as well as of the relevant medicinal drugs are being actively developed and validated. Such models define drug concentration time courses and effects (Pharmacokinetics/Pharmacodynamics, PK/PD) and the physiology of interest at different levels of scale, ranging from molecules (*e.g.*, Roy and Roy, 2010), molecular and gene networks (*e.g.*, Le Novère, 2015), cells (*e.g.*, Bächler *et al.*, 2014), organs (*e.g.*, Cox *et al.*, 2009), up to body compartments (*e.g.*, Balazki *et al.*, 2018) and the whole body (*e.g.*, Hester *et al.*, 2011).

1.1 Motivation

One of the main enablers to perform an ISCT is the availability of a finite *population of virtual patients*, *i.e.*, computational models able to predict (via *simulation*) relevant clinical measurements (those needed to assess efficacy/safety of the therapy, *i.e.*, drug, treatment, or device, under trial) from time courses of clinical actions (such as drug administrations, see, *e.g.*, FDA, 2018; EMA, 2019). For an ISCT to provide *compelling*

evidence of the safety/efficacy of a therapy and to support its design and revision, such population must be *complete*, *i.e.*, representative of the *entire spectrum* of behaviours deemed of interest, from both physiology and drug PK/PD points of view.

Virtual Patients (VPs) are typically derived by *parameterising* quantitative mechanistic VPH models, which in turn are defined by encoding *qualitative* knowledge of the human physiology of interest (*e.g.*, from the literature or pathways databases like KEGG, Kanehisa *et al.*, 2017 or Reactome, Fabregat *et al.*, 2018) as well as PK/PD of pharmaceutical compounds (*e.g.*, Lippert *et al.*, 2019) into mathematical systems such as, *e.g.*, Ordinary Differential Equations (ODEs) or difference equations (see, *e.g.*, Bartocci and Lió, 2016; Irurzun-Arana *et al.*, 2017). Indeed, it is by means of *parameters* (such as stoichiometric constants, rates, or other patient-specific quantities) that such models take into account inter-subject variabilities, as different parameter assignments yield different model trajectories, also in terms of reactions to drug administrations.

1.2 State of the art in computing populations of VPs

Different approaches have been proposed to compute a population of VPs for quantitative VPH models. Such approaches greatly differ depending on whether the given model is *identifiable* or *non-identifiable*.

For *identifiable* models, a *complete* population of VPs can be computed by fitting the models against a set of *in vivo* measurements deemed representative of the *entire spectrum* of behaviours of interest. As an example, the Physiologically-based Pharmacokinetics (PBPK) simulator in (Lippert *et al.*, 2019) provides a large set of VPs compliant with PBPK regulations from EMA, FDA, EFSA, or EPA. Also, in (Kovatchev *et al.*, 2009) a VPH model is described, and a population of 300 VPs is provided for it, representing 100 adults, 100 adolescents, and 100 children. Such VPs have been approved by FDA as a substitute for pre-clinical animal testing of new treatment strategies for Type 1 Diabetes Mellitus. The above models enjoy a very important property: all their parameters describe physiological characteristics, have known ranges of values, and can be reliably estimated through *in-vivo* or *in-vitro* measurements.

The situation becomes more intricate for *non-identifiable* models, for which, to our knowledge, no approach is available to compute *complete* populations of VPs. In fact, although for such models parameter estimation can still be used (*e.g.*, Teutonico *et al.*, 2015; Allen *et al.*, 2016; Rieger *et al.*, 2018; Schmiester *et al.*, 2019; Wang *et al.*, 2020 and citations thereof) to find cases (*counterexamples*) where the therapy under assessment is unsafe/ineffective, the resulting population of VPs is *not* guaranteed to be complete, no matter how large or representative is the input dataset used for fitting. This is because, due to model non-identifiability, there could be other (possibly very different) parameter assignments (not selected through fitting) still matching experimental data, but leading to different model behaviours under the new therapy.

In other words, model non-identifiability hinders the possibility to have a *comprehensive picture* of the cases where the therapy succeeds or fails. As a result, although being based on solid *scientific principles* (*e.g.*, biochemical reactions), thereby satisfying one of the qualification requirements for ISCT (*e.g.*, FDA, 2018; EMA, 2019), it is hard to use non-identifiable models to verify safety/efficacy of a therapy. This is why identifiability is a key test in, *e.g.*, FDA or EMA PBPK guidelines.

In the literature, *qualitative* VPH models have also been considered, for example logic-based models (*e.g.*, Wang *et al.*, 2012; Bloomingdale *et al.*, 2018). Their aim is to predict sequences of Boolean-valued (low *vs.* high) expression levels rather than the time course of the biological quantities of interest. In qualitative models, non-identifiability can somewhat be overcome by modelling lack of knowledge about reaction rates through an asynchronous update schema for their Boolean-valued variables. Complete populations of VPs can then be generated by using finite state model checking techniques to look for *attractors* (*e.g.*, Zheng *et al.*, 2013; Khan *et al.*, 2017; Razzaq *et al.*, 2018 and citations thereof). Unfortunately, this approach cannot be used for quantitative models (like those defined through ODEs or difference equations, our main focus here) defining real-valued (rather than Boolean-valued) concentrations of compounds, where, in general the state space is infinite.

We finally argue that the above problem stemming from non-identifiability also arises in other areas. For example, models used in machine learning (*e.g.*, neural networks) are typically non-identifiable, and it is well known that, notwithstanding how large is the training dataset, it is possible to find (plausible) input data leading to wrong classifications (*e.g.*, Eykholt *et al.*, 2018). Not surprisingly, similarly to ISCT, this is the main obstacle in qualifying machine learning–based approaches within safety-critical (*i.e.*, *high impact* regulatory purpose) applications such as autonomous driving (*e.g.*, Jenn *et al.*, 2020).

The above considerations motivate the main goal of this paper: to develop methods and software that (possibly building on parameter estimation against *in vivo* data) can compute a finite set of *physiologically meaningful*, pairwise *distinguishable* VPs, which are *representative* of the entire spectrum of behaviours defined by the given (possibly non-identifiable) quantitative VPH model (*completeness*).

1.3 Contributions

In this paper we present methods and software to compute populations of VPs for (possibly non-identifiable) quantitative VPH models. We focus on the typical case of models that, due to their complexity, cannot be analysed symbolically, but need to be numerically simulated (*e.g.*, Hucka *et al.*, 2003; Maggioli *et al.*, 2019), and show the effectiveness of our methods on a non-identifiable model of the Hypothalamic-Pituitary-Gonadal (HPG) axis defined in terms of 33 highly non-linear stiff ODEs.

Our populations satisfy three important properties: *completeness*, *pairwise distinguishability*, and *stratifiedness*.

Completeness means that our populations show *all* model behaviours deemed of interest (*e.g.*, *physiologically meaningful*), even when such a full set of behaviours is *unknown* at model design time (this is typical in large non-identifiable, over-parameterised VPH models, see below). For example, the population we computed in our case study comprises as many as 4 830 264 VPs.

Pairwise distinguishability means that no model behaviour (aka *phenotype*) is *over-represented* in our population: any two VPs behave differently (according to some used-defined notions of behavioural distinguishability) in at least one scenario (*e.g.*, input pattern) supported by the model. This avoids waste of computation during an ISCT.

Stratifiedness means that our populations are organised in levels, (strata), each one showing the *entire spectrum* of behaviours under *different distinguishability criteria*. For example, in our case study we stratified our 4 830 264 VPs into 7 sub-populations, each one comprising a number of VPs ranging from 2 million to just 1. Since each sub-population alone is representative of the entire spectrum of model behaviours (of course at different granularity), proper trade-offs can be sought, when designing an ISCT, between the needed behavioural granularity and the budgeted computational effort.

Our *any-time* algorithm, based on *global search* guided by *statistical model checking*, *intelligently explores* the (typically huge) model parameter space, collects those parameter assignments showing a *physiologically meaningful* behaviour (*i.e.*, VPs), and organises them into strata, while guaranteeing a statistically-sound form of *graceful degradation*.

Note that, in many non-identifiable models (like our case-study HPG axis model), most parameter assignments *might not* actually represent

VPs, as, upon simulation, their associated model trajectories show-up to be physiologically meaningless or, anyway, out of interest. This is due to, *e.g.*, over-parameterisation, presence of parameters whose values are not measurable through clinical assays (*e.g.*, reaction rates), presence of unknown (hence, not modelled) interdependency constraints among parameters, and use of parameters to define not-well-understood physiological mechanisms. To find parameter assignments yielding physiologically meaningful model behaviours and different phenotypes is thus computationally very hard, and naïve exploration or sampling of the parameter space could be hopeless.

In order to automatically recognise physiologically meaningful model behaviours (and thus parameter assignments defining VPs), our approach envisions the *elicitation* and *formalisation* of *background biological and medical knowledge* (possibly *also* coming from available data). Our approach is fully independent of how such knowledge is formalised, as long as we can define a criterion that, given a parameter assignment (a candidate VP), decides whether the resulting model trajectory is physiologically meaningful or not.

In our case study, we rely on background knowledge available in terms of known assignments to the model parameters (computed via parameter estimation against clinical data, hence defining *reference* VPs), bounds for model parameters and biological species, and on physiological meaningfulness criteria which ask for (loose) *qualitative similarity* of the model behaviours under a candidate VP with respect to those entailed by some *reference* VP. Such criteria are applicable to a wide class of models, *e.g.*, those defining hormonal signalling networks.

2 Methods

Below we define our framework (Section 2.1) and methodology (Section 2.2) to generate complete stratified populations of pairwise distinguishable VPs.

2.1 Formal framework

VPH models. We adopt a very general approach to define VPH models and view them as parametric input-output dynamical systems. This general definition is standard in signals and systems (see, *e.g.*, Sontag, 1998), especially when, as in the case of physiological models, the system internal state is not accessible, and only selected outputs (*system observables*) can be measured.

Our definition (for a formal statement see Definition 1 in Section SM1.1.1 of the supplementary material) accounts for both *continuous-* as well as *discrete-time* models (*e.g.*, those defined by means of ODEs and difference equations, respectively). Namely, model *inputs* are *time functions* $\mathbf{u}$ defining the time course of exogenous inputs (*e.g.*, drug administrations). Our models are *parametric*, in that their *observation function* $\mathbf{y}(\mathbf{u}, \lambda)$, defining the values $\mathbf{y}(t; \mathbf{u}, \lambda)$ of the system observables at any time point t, depends on *both* the input time function $\mathbf{u}$ *and* the values λ for the system *parameters*, chosen within the model parameter space Λ.

For physical reasons, we require that our VPH models are *strictly causal*, *i.e.*, their observation function up to any time point depends only on *past* inputs. Also, given the presence of parameters, we focus on deterministic systems, in that parameters embody any initial condition which the system output might depend on.

Virtual patients, phenotypes, populations. As anticipated in Section 1, not all assignments to a VPH model parameters yield behaviours of interest. Many might even yield physiologically meaningless behaviours. Conversely, due to, *e.g.*, system over-parametrisation or non-identifiability, multiple parameter assignments may yield (almost) indistinguishable behaviours (*i.e.*, their associated observation functions are very similar on all inputs). Such indistinguishable VPs would increase the computational efforts needed to carry out an ISCT on the entire population, without bringing any advantage in terms of representativeness of the trial.

For generality, our forthcoming definitions rely on user-provided Boolean function φ and equivalence relation $\sim$. Boolean function φ defines the conditions to be met by any parameter $\lambda \in \Lambda$ for the associated model behaviours to be considered of interest, for example *physiologically meaningful* (in which case, λ has to be regarded as a VP). Equivalence relation $\sim$ on the set of VPs defines when two VPs shall be considered having *indistinguishable behaviour* (*i.e.*, showing the same *phenotype*): for any two VPs λ and λ', $\lambda \sim \lambda'$ means that the two VPs show the same phenotype.

With respect to given φ and $\sim$ for a VPH model $\mathcal{S}$, we define the following concepts (for a formal statement see Definition 2 in Section SM1.1.2): (a) the *population* $\hat{\Lambda}$ of VPs for $\mathcal{S}$ is the set of parameter assignments $\lambda \in \Lambda$ for which $\varphi(\lambda)$ is true; (b) the *phenotype* of VP λ is the equivalence class of λ with respect to $\sim$ (notation: $[\lambda]_\sim$); (c) the *phenotype space* $\hat{\Lambda}/\sim$ of $\hat{\Lambda}$ is the quotient set of $\hat{\Lambda}$ with respect to $\sim$, *i.e.*, the set of all-different phenotypes of VPs in $\hat{\Lambda}$; (d) an *All-Different Phenotype Population (APP) of VPs* is any subset $\hat{\Lambda}^\sim$ of $\hat{\Lambda}$ such that no two VPs λ, λ' exist in $\hat{\Lambda}^\sim$ having the same phenotype. Also, an APP $\hat{\Lambda}^\sim$ is said a *Complete APP (CAPP)* if it contains a representative of *all phenotypes* in the phenotype space of $\hat{\Lambda}$.

Clearly, the definition of both function φ and relation $\sim$ depends on the VPH model at hand, and has to be made starting from expert knowledge. Also, in the typical case of models subject to external inputs (*e.g.*, drug administrations), both φ and $\sim$ might need to be defined on model behaviours under *different input functions*. This allows the expert to define meaningfulness and phenotypes of candidate VPs also in terms of their reactions under different sequences of drug administrations (where such reactions are dictated by the PK/PD model equations).

Note that, when $\sim$ is $\mathbf{1}$ (*i.e.*, the equivalence relation defining a distinct class per VP $\lambda \in \hat{\Lambda}$), we have $\hat{\Lambda}/\mathbf{1} = \hat{\Lambda}$. Hence, the entire population of VPs ($\hat{\Lambda}$) can always be regarded as a CAPP.

In Section 3 we give a widely-applicable definition for φ and $\sim$ based on *qualitative similarity* of the model evolutions associated to different parameters.

2.2 Computing complete populations of VPs

Given a VPH model with parameter space Λ, a Boolean function φ and an equivalence relation $\sim$ as in Section 2.1, our goal is to compute a CAPP with respect to φ and $\sim$.

In this paper we focus on cases where the definition of the VPH model, function φ, and the computation of the phenotype $[\lambda]_\sim$ of a VP λ are too complex for set $\hat{\Lambda}^\sim$ to be computed analytically and/or symbolically in closed form. For such complex scenarios, deciding whether $\varphi(\lambda) =$ true or not for any given $\lambda \in \Lambda$ (hence, whether λ represents a VP or not) and, in the affirmative case, computing its phenotype $[\lambda]_\sim$ involves a *numerical simulation* of the VPH model and the subsequent analysis of the resulting model trajectories under different inputs. Also, knowing that $\varphi(\lambda) =$ true for some $\lambda \in \Lambda$ does not allow us to infer (without additional simulations) whether $\varphi(\lambda') =$ true for other parameters $\lambda' \in \Lambda$, let alone their phenotypes.

In order to cope with such a general setting, we adopt a search-based approach that *explores* the model parameter space Λ looking for parameters $\lambda \in \Lambda$ such that $\varphi(\lambda) =$ true and belonging to all-different equivalence classes of $\sim$. This calls for VPH models whose parameter space Λ is finite or can be *finitised* by the user, *e.g.*, into a bounded interval of $\mathbb{N}^k$, $k > 0$. Such finitisation can often be performed by exploiting knowledge about, *e.g.*, physiological bounds to the parameter values and *model locality*

assumptions (*i.e.*, minor changes to the value of a parameter yield minor changes in the resulting model behaviours).

Nevertheless, even when Λ is finite, an exhaustive exploration is practically infeasible unless Λ is very small. Unfortunately, this is not the case for complex VPH models: for example, the size of the (finitised) parameter space of our case-study model is 10^{76}, which makes an exhaustive search clearly out of reach (let alone the fact that computing $\varphi(\lambda)$ for each λ takes seconds of simulation time).

To overcome these obstacles, our search (Section 2.2.1) is an *any-time algorithm* relying on Statistical Model Checking (SMC) and hypothesis testing to guarantee proper statistically-sound *graceful degradation*.

2.2.1 The algorithm

Our algorithm is an *any-time* procedure which builds on the SMC and hypothesis testing methods initially presented in (Grosu and Smolka, 2005) and extended in (Tronci *et al.*, 2014).

Core algorithm. Given a VPH model $\mathcal{S}$ having finite (although too large for an exhaustive exploration) parameter space Λ, plus function φ and equivalence relation $\sim$, our algorithm implements a *one-sided error* procedure to compute a CAPP $\hat{\Lambda}^{\sim}$ for $\mathcal{S}$ with respect to $\sim$. The algorithm randomly samples the parameter space Λ (according to a user-defined *sampling policy*), and iteratively adds to the current $\hat{\Lambda}^{\sim}$ (initialised to $\emptyset$) those parameters λ that represent VPs (*i.e.*, $\varphi(\lambda) = \text{true}$) *and* show a phenotype different than all those already represented in $\hat{\Lambda}^{\sim}$.

The algorithm can be interrupted at *any time* and provides a form of *graceful degradation*: after each sample, the algorithm computes an upper bound $\varepsilon \in (0, 1]$ to the probability that further sampling would produce VPs of unseen phenotypes (*error margin*). This fact would prove that the current APP is not indeed a CAPP. When the achieved value for ε reaches a sufficiently-small (target) threshold, the user can decide to stop the algorithm and get the APP computed so far.

The computed value for ε is a function of the number of consecutive failed attempts N that the algorithm is experiencing in discovering VPs of new phenotypes. Clearly, being based on sampling, our algorithm can commit an error in computing the error margin ε (*i.e.*, it could return a value *lower* than a true upper bound). However, by exploiting statistical hypothesis testing methods, given *any* user-requested value $\delta \in (0, 1)$ (*confidence ratio*), our algorithm ensures (see below and Theorem 1 in Section SM1.2.1) that the probability of such an error is at most δ.

Sampling policy. In order to be effective in discovering VPs of new phenotypes, the employed sampling policy may embody proper *domain expert knowledge* and *structural knowledge* about the VPH model, for example: interdependency constraints among components of the parameter values (very common in over-parameterised models), or sensitivity information of model behaviours with respect to parameter values. Also, the sampling policy can be *refined* and *improved* during the search to embed new knowledge, *e.g.*, about the newly discovered VPs. In Section 3.4 we will outline a sampling policy for our case-study model (but widely applicable in general), which exploits the above flexibility.

Parallel computation. Our algorithm takes advantage of a parallel High Performance Computing (HPC) infrastructure. The parameter space Λ is split upfront into k slices $\Lambda_1, \ldots, \Lambda_k$, and k independent instances of our core algorithm can be run in parallel, where instance i ($i \in [1, k]$) draws samples from Λ_i to build population $\hat{\Lambda}_i^{\sim}$. When $\hat{\Lambda}_i^{\sim}$ is computed for all slices, a final population $\hat{\Lambda}^{\sim}$ is produced by taking the union of the phenotype spaces of all $\hat{\Lambda}_i^{\sim}$ and by choosing one representative VP from each equivalence class. To take load balancing into account, the overall number of parallel processes can be much higher than the number of slices (k). An orchestrator can then dynamically assign such processes to the exploration of each slice, in order to keep the values of ε balanced. This approach to parallelism and load balancing is very effective (see, *e.g.*, Mancini *et al.*, 2016) and avoids overhead due to inter-process communication (as that experienced in, *e.g.*, Mancini *et al.*, 2015).

function slice_APPs($\mathcal{S}, \Lambda_i, \varphi, \sim_1, \ldots, \sim_L, \delta$)
  $\hat{\Lambda}^{\sim}{}_{\sim_1 i}, \ldots, \hat{\Lambda}^{\sim}{}_{\sim_L i} \leftarrow \emptyset$; $N_1, \ldots, N_L \leftarrow 0$;
  while not interrupted
    $\lambda \leftarrow$ new sample from Λ_i according to sampling policy;
    foreach $l \in [1, L]$ **do**
      if $\varphi(\lambda)$=true and $[\lambda]_{\sim_l}$ unknown in $\hat{\Lambda}_i^{\sim_l}$ **then** add λ to $\hat{\Lambda}_i^{\sim_l}$; $N_l \leftarrow 0$; $\varepsilon_l \leftarrow 1$;
      else N_l++; $\varepsilon_l \leftarrow 1 - \delta^{1/N_l}$;
      output $(\sim_l, \hat{\Lambda}_i^{\sim_l}, \varepsilon_l)$;
    if sampling policy to be revised **then** revise policy; $N_1, \ldots, N_L \leftarrow 0$;
end

Fig. 1. A parallel branch of our any-time algorithm to compute stratified APPs.

Simultaneous computation of stratified APPs. Our algorithm can work with multiple equivalence relations $\sim_1, \ldots, \sim_L$, defining different behavioural indistinguishability (*i.e.*, same phenotype) criteria, *e.g.*, at different levels of abstraction. When it makes sense to use the same policy to sample the VPH model parameter space Λ for all the $\sim_l$ ($l \in [1, L]$), then the L CAPPs can be computed simultaneously using the *same sequence* of random samples. In Section 3 we will exploit this possibility to compute a hierarchy of stratified CAPPs for our case-study VPH model.

Complete algorithm and main result. Let $\Lambda_1, \ldots, \Lambda_k$ be a partitioning of the finite (or finitised) parameter space Λ of our VPH model $\mathcal{S}$ into $k > 0$ slices. Our overall algorithm runs in parallel k instances of the algorithm in Figure 1, where instance $i \in [1, k]$ runs on slice Λ_i of Λ and computes $L > 0$ APPs, one for each given equivalence relation $\sim_l$ ($l \in [1, L]$) on the population of VPs entailed by the given function φ. During computation, each parallel branch (Figure 1) outputs a stream of tuples of the form $(\sim_l, \hat{\Lambda}_i^{\sim_l}, \varepsilon_l)$ (one after each sample and for each equivalence relation $\sim_l$). Each such tuple states that (for a formal statement see Theorem 1 in Section SM1.2.1), with statistical confidence $(1 - \delta)$, the probability that further sampling within Λ_i will disprove that $\hat{\Lambda}_i^{\sim_l}$ is a CAPP of Λ_i with respect to $\sim_l$ is $< \varepsilon_l$. The algorithm in Figure 1 includes a periodic revision of the sampling policy in order to exploit the new acquired knowledge (of course at the price of resetting all counters $N_l, l \in [1, L]$).

3 Computing complete stratified populations for a VPH model of the HPG axis

In this section we show how we instantiated the general methodology described in Section 2 to a complex state-of-the-art VPH model of the HPG axis (called GynCycle) in order to compute a *stratified* set of CAPPs. We argue that our approach is based on general concepts applicable to a wide class of VPH models, *e.g.*, those defining hormonal signalling networks.

3.1 The GynCycle model

GynCycle (Röblitz *et al.*, 2013) is a VPH model of the human female HPG axis with a special focus on the interactions and feedback mechanisms at different stages of the menstrual cycle. The model (see Section SM2.1 for more details) defines, by means of parametric highly non-linear ODEs, the dynamics of 33 biological *species* (mostly hormones) having a role in the menstrual cycle (*e.g.*, GnRH, FSH, LH, E2, P4 among the others) and the PK/PD of two pharmaceutical compounds. In particular, model inputs encode administrations of GnRH analogues that alter the menstrual cycle.

We formalised our GynCycle model as a dynamical system $\mathcal{S}$ (Section 2.1) as follows.

Time span. Due to the model complexity, GynCycle evolutions need to be computed by *numerical simulation*. This results in both input and observation functions being *bounded-horizon* sequences of samples *evenly spaced in time*. To obtain robust results, we computed physiological meaningfulness metrics (Section 3.2) and phenotypes (Section 3.3) across 120 days (*i.e.*, roughly 4 menstrual cycles), after ignoring the first 3 cycles (to get rid of any *transient model behaviours*, with this value being established by preliminary experiments). The *time quantum* between samples was set to 14.4 minutes (*i.e.*, 100 samples per day) to account for the physiological time scales of the modelled signalling pathways. Hence, input and observation functions are encoded as sequences of $h = 12\,000$ samples, one every 14.4 minutes.

Parameter space. The model counts 76 real-valued *patient-specific parameters* (*e.g.*, hormone decay rates, reaction rates, stimulatory and inhibitory effects) with known bounds (Röblitz *et al.*, 2013). By preliminary experiments we assessed that a change of parameter values of $<$10% yields very small changes in the resulting model trajectories (model locality). Hence, by discretising the interval for each parameter into 10 values, we produced a finitised parameter space Λ of size 10^{76}. Although finite, this size is still too large to be explored exhaustively. However, thanks to our informed sampling policy (Section 3.4), we were able to compute large APPs proved complete with a high statistical confidence (95%) and a small error margin (as low as 5×10^{-5}).

Model inputs. Model inputs define doses for each of the two supported pharmaceutical compounds. Thus, an input time function defines a time sequence of doses administered for each of the two compounds.

Model outputs. Model outputs are non-negative real values for the $n \in \mathbb{N}_+$ model observables. In Section 3.5 we experiment with $n = 4$ observables, namely: LH, FSH, E2, P4, which are the hormones typically measured in a clinical setting, and for which we have retrospective data (Section 3.5.3).

3.2 Physiological meaningfulness

In (Röblitz *et al.*, 2013), GynCycle has been fitted against a database (courtesy of Pfizer) comprising 20–25 measures for 4 observed hormones (E2, P4, FSH and LH) on 12 healthy women, totalling more than 1000 measurements. This activity produced a parameter assignment $\lambda^{(0)} \in \Lambda$ which entails model behaviours *averaging* those of such 12 patients (see Section SM2.2).

In hormonal signalling pathways like those in GynCycle, all healthy humans show the *same qualitative time course* of such hormones. Hence, $\lambda^{(0)}$ defines a VP that we can (and do) regard as a *reference* VP. Thus, we defined function φ (which encodes the physiological meaningfulness criteria that must be satisfied by a parameter assignment λ for it to be considered a VP, see Section 2.1) asking for (loose) *qualitative similarity* between the model observation functions under λ and those under $\lambda^{(0)}$. Namely, we proceed as outlined in the following sections.

Representative portfolio of input functions. In order to derive VPs whose behaviour is meaningful also when drugs are administered, we defined a *representative portfolio* $\mathbf{U}$ of 5 different input functions. Beyond the no-drug input (under which the GynCycle observation function must represent a healthy natural menstrual cycle), we considered two standard treatment strategies, consisting of daily administrations of two different doses for each of the two pharmaceutical compounds supported by the model (see Section SM2.2.1).

Physiological meaningfulness as qualitative similarity. Our function φ returns *true* on $\lambda \in \Lambda$ (thus declaring λ to be a VP), if and only if the model observation functions under λ, when subject to *each* of the input functions in $\mathbf{U}$, have values always within certain *physiological bounds*, and can be (jointly) *time-scaled* and/or *time-shifted* (up to a certain limit) so to satisfy certain *qualitative similarity metrics*, when compared to the observation functions entailed by the reference VP $\lambda^{(0)}$ under the *same* input. Time shifting and scaling allow us to deal with time-alignment issues and different menstrual cycle durations, respectively.

The qualitative similarity metrics we exploited are standard (discrete-time) *signal processing metrics* (see, *e.g.*, Vaseghi, 2009): the *Normalised Zero-Lag Cross-Correlation (NZC)* and the *Normalised Energy Difference (NED)*, which we require to be, respectively, above and below certain thresholds. In our experiments, we set such thresholds to 70% and 80%, respectively. We also set limits for time-scaling and time-shifting to $\pm$10% and 35 days, respectively. Such values (defined after preliminary experiments) are generous enough to allow us to accept model behaviours quite different from those entailed by the reference VP, but still appearing physiologically meaningful to a visual inspection.

The intuition behind and the formal definitions of our metrics, as well as technical details on how $\varphi(\lambda)$ is actually computed (for any given $\lambda \in \Lambda$) are reported in Section SM2.2.2. Here, we just point out that such computations are quite heavy. In particular, GynCycle must be numerically simulated under each candidate parameter λ and each input function $\mathbf{u} \in \mathbf{U}$, in order to retrieve the observation function $\mathbf{y}(\mathbf{u}, \lambda)$. Also, time-scaling and time-shifting issues must be evaluated before computing our similarity metrics between $\mathbf{y}(\mathbf{u}, \lambda)$ and $\mathbf{y}(\mathbf{u}, \lambda^{(0)})$. To cope with such issues efficiently, our approach envisions the solving of a constraint satisfaction problem to enumerate all possible *peak alignments* between the two observation functions, and the use of algorithms to compute NZC and NED between (the time-scaled and time-shifted) $\mathbf{y}(\mathbf{u}, \lambda)$ and $\mathbf{y}(\mathbf{u}, \lambda^{(0)})$, for each $\mathbf{u} \in \mathbf{U}$.

3.3 Stratified phenotypes

Our definition of behavioural indistinguishability (*i.e.*, same-phenotype equivalence relation) of different VPs follows an approach consistent to the one we used to decide physiological meaningfulness. However, in this case, similarity is *quantitatively* evaluated between the observation functions of each pair of VPs (*i.e.*, parameters that, by satisfying function φ in Section 3.2, already satisfy the qualitative similarity metrics thresholds against the reference VP $\lambda^{(0)}$).

To compare two observation functions available in the form of discrete sequences of real-valued samples evenly spaced in time (as is our case), we compare the coefficients of their Discrete Fourier Transforms (DFTs) (see, *e.g.*, Vaseghi, 2009). In particular, to define behavioural indistinguishability among VPs, we use an equivalence relation $\sim_\psi$, parametric in $\psi \in \mathbb{R}_+$ (the *quantisation factor*). Two VPs $\lambda^{(1)}$ and $\lambda^{(2)}$ belong to the same equivalence class (*i.e.*, $\lambda^{(1)} \sim_\psi \lambda^{(2)}$) if and only if the DFT coefficients of their associated VPH model observation functions (for all observables and for all input functions $\mathbf{u} \in \mathbf{U}$) belong to the same quantum (for a formal statement see Definition 4 in Section SM2.3). The size of quanta for DFT coefficients is inversely proportional to both ψ and the energy of the observation function of each model observable $i \in [1, n]$ under the distinguished parameter assignment $\lambda^{(0)}$ ($\left|\left|\mathbf{y}_i(\mathbf{u}, \lambda^{(0)})\right|\right|^2$), which acts as a *normalising factor*. This is important, because the different model observables may assume values in very different ranges. In our experiments (Section 3.5) $\lambda^{(0)}$ is the GynCycle reference VP.

Our definition of $\sim_\psi$ implies (see Remark 1 in Section SM2.3) that ψ is an upper bound to the *NED* shown by the observation functions of *any two* VPs $\lambda^{(1)}$ and $\lambda^{(2)}$ such that $\lambda^{(1)} \sim_\psi \lambda^{(2)}$, for any model observable $i \in [1, n]$ and input function $\mathbf{u} \in \mathbf{U}$. Thus, by considering L increasing values for ψ: $\psi_1 < \cdots < \psi_L$ ($L \in \mathbb{N}_+$), we define L equivalence relations $\sim_{\psi_1}, \ldots, \sim_{\psi_L}$ that group VPs in larger behavioural indistinguishability classes as their associated quantisation factor increases (*stratified phenotypes*). In our experiments (Section 3.5), we choose $L = 7$

and an increasing set of 7 values for ψ (see Table 1), where ψ_L is such to place all generated VPs into a *single* equivalence class.

Indeed, value ψ turns out to be a *very loose* upper bound for the NED between VPs belonging to the same equivalence class. This is because it does not take into account the fact that all our VPs are known to satisfy the physiological meaningfulness criteria of Section 3.2 (qualitative similarity with respect to the behaviour of the VPH model under parameter $\lambda^{(0)}$). In particular, since such criteria depend on optimal time-shifts and time-stretches sought *for each single* VP, our bound to the NED cannot exploit such knowledge and needs to stick to the worst-case. To this end, in our experimental analysis, we also compute, by means of auxiliary hypothesis testing–based SMC tasks (along the lines of our main algorithm of Section 2.2.1, with error margin 1% and confidence ratio 5%), the *actual* maximum NED between VPs belonging to the same equivalence class of each stratum (see Table 1).

3.4 Sampling policy and parallel computation

Like many VPH models, GynCycle is organised in several components, one for each of the modelled hormones. Changing the values of the elements of the parameter vector occurring in a few components typically changes the overall model dynamics only partially.

This key observation is at the heart of our sampling policy. In order to draw, with high probability, a parameter assignment that proves to be a VP, we exploit the knowledge acquired in the past iterations, in terms of the parameter assignments that already proved to be VPs. Namely, let $\hat{\Lambda}_{\text{current}}$ be the set of VPs already discovered (*population of known VPs*). Our sampling policy draws a random parameter λ by changing uniformly at random the elements occurring in $p \in \mathbb{N}_+$ model components (chosen uniformly at random) from a parameter $\hat{\lambda}$ chosen uniformly at random from $\hat{\Lambda}_{\text{current}}$ (if $\hat{\Lambda}_{\text{current}}$ is empty, then $\hat{\lambda} = \lambda^{(0)}$). Value of p is drawn from a Zipf's distribution (*i.e.*, $p \sim ap^{-b}$, where a is a normalisation factor), in order to draw with high probability small values. In our experiments we set b to 3 so that the expected value for p is about 1.11.

The sampling policy is periodically revised by updating $\hat{\Lambda}_{\text{current}}$ with the new discovered VPs. However, in order to avoid too frequent policy updates (which would resort in an immediate reset of the consecutive failure counters, see Section 2.2.1), set $\hat{\Lambda}_{\text{current}}$ is updated only every a given number N of samples. In our experiments we chose N such that experiencing N consecutive failures to find a new VP (regardless of its phenotype) would allow us to conclude, with statistical confidence $(1 - \delta) = 95\%$, that the probability that additional VPs will be found by further sampling is less than $\varepsilon = 1 - \delta^{1/N} = 5 \times 10^{-5} = 0.005\%$. This results in $N = 59\,914$.

For the above sampling policy to work on top of a slicing of the parameter space Λ to be processed in parallel, it is enough to ensure that $\lambda^{(0)}$ belongs to all slices. This was done by defining our (initially continuous) parameter space finitisation as a grid having $\lambda^{(0)}$ as one of its vertices, and by defining the k slices by bisecting Λ on values $\lambda^{(0)}_{i_1}, \ldots, \lambda^{(0)}_{i_r}$ for any subset of coordinates $i_1, \ldots, i_r$ within $[1, 76]$, thus defining $k = 2^r$ slices $\Lambda_1, \ldots, \Lambda_k$ all containing $\lambda^{(0)}$. In our experiments we chose $r = 7$ random coordinates, hence $k = 128$.

3.5 Experimental results

Here we present our results on GynCycle. In Section 3.5.1 we show the APPs we computed, in Section 3.5.2 we analyse the behaviour of our sampling policy, and in Section 3.5.3 we perform a qualitative and quantitative evaluation of the representativeness of our populations with respect to retrospective clinical data (86 medical cases courtesy of Hannover Medical School, University Hospital of Lausanne, and Pfizer).

id	ψ	APP size	error margin (ε)			max NED
			min	avg	max	
7	16 200	1	5×10^{-5}	5×10^{-5}	5×10^{-5}	163.23%
6	8100	104	5×10^{-5}	5×10^{-5}	5×10^{-5}	144.36%
5	4050	3862	5×10^{-5}	5×10^{-5}	5×10^{-5}	106.74%
4	2700	43 941	5×10^{-5}	6.75×10^{-3}	4.51×10^{-1}	84.70%
3	1800	251 239	5.09×10^{-4}	2.36×10^{-2}	1	59.09%
2	900	2 136 710	3.25×10^{-3}	8.33×10^{-2}	1	48.07%
1	–	4 830 264	9.87×10^{-3}	1.81×10^{-1}	1	–

Table 1. Stratified GynCycle APPs. Statistical confidence: 95%.

3.5.1 Computed APPs

We ran our SMC-based algorithm on a parallel HPC infrastructure (the Marconi cluster at Cineca, Italy) with the settings defined above, in order to compute the stratified APPs as defined in Section 3.3. Confidence ratio δ was set to 0.05.

The computation was stopped after around 60 days. In total, our algorithm sampled 414 245 648 parameters (simulating GynCycle for 7 months on each of them and for each of the input functions in the representative portfolio described in Section 3.2). Overall, 4 830 264 parameters were declared to define VPs.

Table 1 lists the sizes of the 7 computed APPs. The bottom line refers to the entire population of VPs, $\hat{\Lambda}^{\mathbf{1}}$ (which is an APP with respect to equivalence relation $\mathbf{1}$).

We decided to terminate our (any-time) computation when we achieved $\varepsilon = 5 \times 10^{-5}$ for *all slices* on the *top three* strata. This means that (see Section 2.2.1 and Theorem 1 in Section SM1.2.1), with statistical confidence $1 - \delta = 95\%$, the probability that further sampling (in *any* single slice) would disprove that such top three APPs are indeed CAPPs is below the error margin (5×10^{-5}).

As for the other strata, the table reports minimum, maximum and average error margins across the $k = 128$ parallel processes (one per slice) at the time of termination of our any-time computation. Since the exploration of each slice is an independent process, the k error margins for each stratum can be quite different, as the value for ε for a given slice depends on the time when the last VP belonging to *that* slice was generated. Also, when we terminated our experiment, a new VP (of a phenotype *known* to the top three strata) was *just* generated. Hence, the max ε for the population $\hat{\Lambda}^{\mathbf{1}}$ consisting of *all* VPs (bottom line of Table 1) is 1.

Figure 2(a) shows the trajectories of the GynCycle observables under the VPs belonging to the computed APPs for all strata except the extreme two. It can be seen that, despite the number of VPs greatly reduces at higher levels of our stratification, all APPs retain full *representativeness* of the entire spectrum of possible behaviours.

A final note is in order. Although 60 days could appear an unusually-long time for a computation (especially if compared to the time typically needed by classical model fitting tasks), this is a *one-time activity* for the input VPH model, and can be sped-up almost arbitrarily by using a higher number of parallel processes (*e.g.*, using 1280 processes –which is perfectly feasible in today's infrastructure-as-a-service platforms– with groups of 10 processes jointly exploring each of our 128 slices, would have required just 6 days). Indeed, once a population of VPs for a given model has been computed, it can be used to carry-out *multiple* ISCT (*i.e.*, for different treatment strategies or medical devices). Each ISCT can be carried-out on the *most appropriate stratum* of VPs, depending, *e.g.*, on the chosen trade-off between budgeted computational effort and required behavioural granularity of the VPs recruited for the trial. Also, more sophisticated approaches can be exploited, *e.g.*, *iterative deepening* within the stratification of phenotypes (guided by simulation results) searching for a VP showing a failure of the candidate treatment or medical device (a *counter-example*, see, *e.g.*, Mancini *et al.*, 2013).

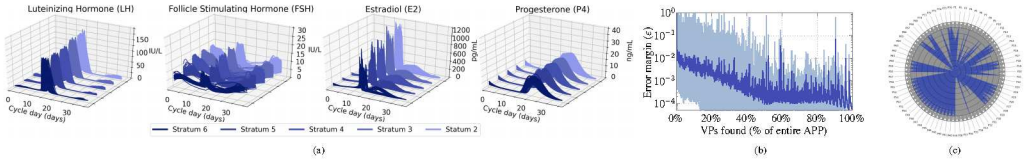

Fig. 2. (a) Time evolutions for the GynCycle observables under the VPs belonging to the computed stratified APPs. (b) Average error margin (ε) reached during parallel computations (stratum 5). (c) Parameter space exploration.

3.5.2 Sampling policy behaviour

Our *informed* sampling policy was able, on average, to find (within our 128 slices) an admissible VP every 86 attempts (average success rate: 1.17%). This is to be compared to a *uniform (non-informed) sampling policy*, which was *unable* to discover a single VP after *50 million attempts*.

Figure 2(b) shows the error margin achieved by our informed sampling policy during generation of $\hat{\Lambda}^{\sim 4050}$, *i.e.*, the APP associated to the smallest value of ψ (see Table 1) for which we reached an error margin of 5×10^{-5} for all slices. The plot shows the values for the error margin reached by each of the 128 parallel computations (light curves) when discovering each of its VPs (x axis), thus disproving that the current APP was indeed a CAPP. Values for x have been normalised into percentages of the total number of the VPs discovered by each parallel computation. We note that the average error margin (dark curve) lies for most of the time at values *one order of magnitude higher* than the value we chose to terminate our experiments ($\varepsilon = 5 \times 10^{-5}$, see Table 1). This shows that our informed sampling policy was *always* effective to extract (with probability much higher than 5×10^{-5}) new VPs when (we know that) they actually exist.

Finally, Figure 2(c) shows, as a radar plot, the location of VPs within the GynCycle parameter space. The figure shows one polygon per VP, which connects the chosen values for the 76 parameter vector elements. Interestingly, for some of them, only a few values of their domains actually occur in VPs. Such constraints were *unknown* at the time of model design.

3.5.3 Validation against clinical data

The previous sections show that our computed populations exhibit the properties of *pairwise distinguishability* and *stratifiedness*, as well as that *representativeness* of the spectrum of behaviours is kept among the different strata.

What remains to be shown is that our sampling policy was indeed able to extract VPs representative of the entire space of physiologically meaningful behaviours that our input VPH model is *capable* to represent. Such a full set is of course not known. However, the GynCycle was experimentally shown in (Röblitz *et al.*, 2013) to be expressive enough to correctly represent a wide spectrum of behaviours of healthy women.

Hence, here we compare the behaviours shown by our VPs with respect to retrospective clinical data we got from 86 health records, kindly made available to us by Hannover Medical School (35 patients), University Hospital of Lausanne (39 patients) and Pfizer (12 patients, which were originally used in (Röblitz *et al.*, 2013) to compute the reference GynCycle VP). In each dataset, for each health record we have actual measurements of the blood levels of the 4 model observables (LH, FSH, E2, P4) on a (roughly) daily basis for an entire menstrual cycle (all health records refer to healthy patients subject to no pharmaceutical treatment).

Below we perform both a qualitative and a quantitative assessment of the representativeness of our computed VP population against such datasets.

Qualitative evaluation. Figure 3(a) shows daily blood hormone levels on the 86 health records (box-and-whisker plots) on top of the model observation functions (*i.e.*, the time functions of the 4 model observables) of all VPs in our full APP (*i.e.*, $\hat{\Lambda}^1$ of Table 1, blue curves). Curves as well as data have been aligned on the LH peak (used to estimate the ovulation day), in order to account for different transient periods among our VPs. The figure shows that our VP population is indeed highly representative of the available clinical measurements, and that the qualitative behaviours of our VPs faithfully reflect those of the available data.

Quantitative evaluation. Our approach for a quantitative evaluation of the representativeness of our full APP $\hat{\Lambda}^1$ with respect to the human behaviours occurring in our datasets, has been shaped on the fact that $\hat{\Lambda}^1$ does not define a probability distribution of behaviours. In particular, although $\hat{\Lambda}^1$ might (and indeed does, see Table 1) contain VPs showing similar behaviours (which are then removed from the higher strata of our hierarchy), the number of VPs exhibiting any behaviour has *no relation* with the frequency of that behaviour in the real world, but only depends on the model ODEs and on the definition and usage of parameters within them. This implies that statistical approaches to measure the similarity between our APP and the distribution of behaviours shown in our datasets (*e.g.*, those based on the relative entropy of two probability distributions or the similarity of their momenta) cannot be employed in our case.

To assess the representativeness of our APP with respect to the available datasets, we then proceed at computing a deterministic measure of *coverage*, by assessing the percentage of health records for which there exist a VP in our APP exhibiting a *good-enough fit*. Such measure is defined in terms of a given upper bound of a standard error metric, the Average Normalised Mean Absolute Error (aNMAE).

Full details on how we formalise each health record in our datasets and on how we compute the aNMAE of each VP with respect to it are delayed to Section SM3. Here, we comment on Figure 3(b), which shows the coverage of our three datasets as a function of the aNMAE, as resulting from our analysis. The figure shows that most health records are covered by our population within *small* aNMAE values. Namely, the totality of the Pfizer, Hannover, and Lausanne medical records are covered within aNMAE 15%, 20%, and 35%, respectively. As for the latter dataset, 90% of the cases are actually covered within an aNMAE of just 20%.

4 Conclusions

In this paper we presented methods and software to compute a *complete and stratified population of pairwise distinguishable VPs* for a given quantitative model of the human physiology (plus drugs PK/PD). Our approach is especially designed for complex (*e.g.*, non-linear stiff ODE–based) parametric non-identifiable VPH models that cannot be analysed symbolically or integrated in closed form, but must be numerically simulated. To this end, our algorithm runs a global search on the space of model parametererisations, guided by statistical model checking and hypothesis testing, and exploiting suitable biological and medical knowledge elicited from experts to recognise physiologically meaningful behaviours and different phenotypes, as well as structural knowledge of the model to intelligently drive the search via an informed sampling

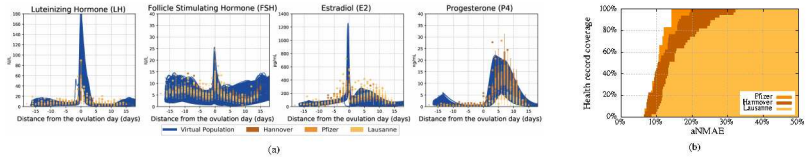

Fig. 3. (a) Qualitative and (b) quantitative validation of our GynCycle population against clinical data.

policy. Our algorithm can be stopped at *any time*, since it continuously provides an upper bound (correct with a user-defined confidence level) to the probability that further computation will discover new phenotypes.

We proved the effectiveness of our algorithm on a state-of-the-art non-identifiable ODE-based VPH model of the female HPG axis, by generating a population of 4 830 264 VPs stratified into 7 levels (at different granularity of behaviours), and assessed its representativeness against 86 retrospective health records.

References

Allen, R. *et al.* (2016). Efficient generation and selection of virtual populations in quantitative systems pharmacology models. *CPT: Pharmacom Sys Pharmacol*, **5**(3).

Avicenna Project (2016). *In silico* clinical trials. `avicenna-isct.org`

Bächler, M. *et al.* (2014). Species-specific differences in follicular antral sizes result from diffusion-based limitations on the thickness of the granulosa cell layer. *Mol Hum Reprod*, **20**(3).

Balazki, P. *et al.* (2018). A quantitative systems pharmacology kidney model of diabetes associated renal hyperfiltration and the effects of sglt inhibitors. *CPT: Pharmacom Sys Pharmacol*, **7**(12).

Bartocci, E. and Lió, P. (2016). Computational modeling, formal analysis, and tools for systems biology. *PLoS Comput. Biol.*, **12**(1).

Bloomingdale, P. *et al.* (2018). Boolean network modeling in systems pharmacology. *J Pharmacokin Pharmacodyn*, **45**(1).

Chis, O.-T. *et al.* (2011). Structural identifiability of systems biology models: A critical comparison of methods. *PLoS ONE*, **6**(11).

Cox, L. *et al.* (2009). A mathematical model to evaluate control strategies for mechanical circulatory support. *Artif Organs*, **33**(8).

European Medicines Agency (2019). Reporting of physiologically based pharmacokinetic (PBPK) modelling and simulation. EMA/CHMP/458101/2016. `ema.europa.eu`

Eykholt, K. *et al.* (2018). Robust physical-world attacks on deep learning visual classification. In *IEEE CVPR 2018*.

Fabregat, A. *et al.* (2018). The Reactome pathway knowledgebase. *Nucl Acids Res*, **46**(D1), D649–D655.

Food and Drug Administration (2018). Physiologically based pharmacokinetic analyses – format and content guidance for industry. FDA-2016-D-3969. `fda.gov`

Grosu, R. and Smolka, S. (2005). Monte Carlo model checking. In *TACAS 2005*, *LNCS* 3440. Springer.

Hester, R. *et al.* (2011). Hummod: a modeling environment for the simulation of integrative human physiology. *Front Physiol*, **2**.

Hucka, M. *et al.* (2003). The Systems Biology Markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**(4).

Irurzun-Arana, I. *et al.* (2017). Advanced boolean modeling of biological networks applied to systems pharmacology. *Bioinformatics*, **33**(7).

Jenn, E. *et al.* (2020). Identifying challenges to the certification of machine learning for safety critical systems. In *ERTS 2020*.

Kanehisa, M. *et al.* (2017). Kegg: New perspectives on genomes, pathways, diseases and drugs. *Nucl Acids Res*, **45**(D1).

Khan, F. *et al.* (2017). Unraveling a tumor type-specific regulatory core underlying e2f1-mediated epithelial-mesenchymal transition to predict receptor protein signatures. *Nat Commun*, **8**(1).

Kovatchev, B. *et al.* (2009). In silico preclinical trials: A proof of concept in closed-loop control of type 1 diabetes. *JDST*, **3**(1).

Le Novère, N. (2015). Quantitative and logic modelling of molecular and gene networks. *Nat Rev Genet*, **16**(3).

Lippert, J. *et al.* (2019). Open systems pharmacology community–an open access, open source, open science approach to modeling and simulation in pharmaceutical sciences. *CPT: Pharmacom Sys Pharmacol*, **8**(12).

Maggioli, F. *et al.* (2019). SBML2Modelica: Integrating biochemical models within open-standard simulation ecosystems. *Bioinformatics*.

Mancini, T. *et al.* (2013). System level formal verification via model checking driven simulation. In *CAV 2013*, *LNCS* 8044. Springer.

Mancini, T. *et al.* (2015). Computing biological model parameters by parallel statistical model checking. In *IWBBIO 2015*. Springer.

Mancini, T. *et al.* (2016). SyLVaaS: System level formal verification as a service. *Fundam Inform*, **1–2**.

Pappalardo, F. *et al.* (2019). In silico clinical trials: concepts and early adoptions. *Brief Bioinf*, **20**(5).

Razzaq, M. *et al.* (2018). Computational discovery of dynamic cell line specific boolean networks from multiplex time-course data. *PLoS Comput Biol*, **14**.

Rieger, T. *et al.* (2018). Improving the generation and selection of virtual populations in quantitative systems pharmacology models. *Progr Biophys Mol Biol*, **139**.

Röblitz, S. *et al.* (2013). A mathematical model of the human menstrual cycle for the administration of GnRH analogues. *J Theor Biol*, **321**.

Roy, P. and Roy, K. (2010). Molecular docking and qsar studies of aromatase inhibitor androstenedione derivatives. *JPP*, **62**(12).

Schmiester, L. *et al.* (2019). Efficient parameterization of large-scale dynamic models based on relative measurements. *Bioinformatics*, **36**(2).

Sontag, E. (1998). *Mathematical Control Theory: Deterministic Finite Dimensional Systems (2nd Ed.)*. Springer.

Teutonico, D. *et al.* (2015). Generating virtual patients by multivariate and discrete re-sampling techniques. *Pharm Res*, **32**(10).

Tronci, E. *et al.* (2014). Patient-specific models from inter-patient biological models and clinical records. In *FMCAD 2014*. IEEE.

Vaseghi, S. (2009). *Advanced Digital Signal Processing and Noise Reduction*. Wiley.

Wang, H. *et al.* (2020). Conducting a virtual clinical trial in HER2-negative breast cancer using a quantitative systems pharmacology model with an epigenetic modulator and immune checkpoint inhibitors. *Front Bioeng Biotech*, **8**.

Wang, R.S. *et al.* (2012). Boolean modeling in systems biology: an overview of methodology and applications. *Phys Biol*, **9**(5).

Zheng, D. *et al.* (2013). An efficient algorithm for computing attractors of synchronous and asynchronous boolean networks. *PLoS ONE*, **8**(4).