Retiring Occam's razor in mechanistic modelling: the role of symmetries in model selection

Johannes Borgqvist[†‡§¶] and Sam Palmer[†].

[†]Wolfson Centre for Mathematical Biology, Mathematical Institute, University of Oxford, United Kingdom
[‡] Research fellow at the Wenner-Gren Foundations, Sweden
[§] Junior Research Fellow (JRF) at Linacre College, University of Oxford, United Kingdom
[¶]E-mail: Johannes.Borgqvist@maths.ox.ac.uk

Introduction

Despite the achievements of mathematical modelling in biology there is still a substantial fundamental problem hindering us from theoretically understanding these complex systems. Currently, it is not possible in many situations to capture the underlying mechanisms of the studied system in the theoretical description with a high certainty. This is mainly explained by the shear intricacy of living systems which further complicates the task of gathering enough experimental data as well as interpreting it in order to build a comprehensive body of knowledge of the system at hand. On account of these "knowledge gaps", theoreticians are bound to make certain assumptions when constructing a model which typically falls into one of two categories depending on what the purpose of the model is. Firstly, there are *statistical models* in which the model is a tool for understanding trends in the data (e.g. linear regression) and often the structure of the model is not of interest. Secondly, there are *mechanistic models* (e.g. coupled systems of differential equations) whose structures are essential and here the goal is to build a coherent theoretical description which connects multiple interacting unobservable parts in order to ultimately increase the knowledge of the biological entity at hand. If models from the latter class can be validated by describing an observed property or collected data, then the final objective is to make predictions of future unknown outcomes and the grand obstacle preventing the construction of predictive mechanistic models lies in the validation step.

As has been observed in numerous situations such as tumour growth in the context of cancer[1], there are often multiple different and potentially *mutually exclusive* models that all describe the same data equally well. In terms of mechanistic modelling, this corresponds to the existence of multiple plausible explanations or hypotheses about the function of the studied system as the structure of every model encodes a specific mechanism. From a theoretical perspective, this is formulated as a *model selection problem* in which multiple candidate descriptions are calibrated to experimental observations and then the candidate that best fit the data is selected. In the situation where multiple models fit the data equally well, often the model that is simplest (e.g. smallest meaning fewest components) is selected within statistical and engineering applications. The philosophical underpinnings of this line of reasoning stem form the principle called *Occam's razor* initially formulated as "*do not multiply entities beyond necessity*"[2]. Even in mechanistic modelling, these sorts of statistical methods are used when selecting a theoretical description which is highly problematic as biology is known as the study of complexity and it is not necessarily the case that the simplest explanation is more realistic than other more complicated counterparts. Luckily, these sorts of questions have been addressed with huge success in theoretical physics which can serve as a source of inspiration for biological applications.

The underlying properties of physical systems have been theoretically described by so called *symmetries*. These objects describe conserved properties often referred to as *invariants* and using these it is possible to formulate *conservation laws*[3] capturing the fundamental properties of the studied system in a theoretical formula. Mathematically, symmetries are operators which leave the structure of the objects they transform intact (e.g. rotations of an equilateral triangle with an angle of 180°), and in the context of differential equations the symmetries map a solution curve to another solution curve[4, 5]. Although there are examples of using symmetries to analyse mechanistic models in biology[6], these techniques are not widespread specially not when it comes to guide model selection based on biological properties encoded by the symmetries as oppose to basing it on Occam's razor. An example of a symmetry based model selection with simulated

data studied a class of models composed of *ordinary differential equations* (ODEs) called the Hill equation describing the kinetics of an enzyme and here it was possible to find the correct model in a situation where multiple initially indistinguishable models all fitted the data equally well [7]. Using this study as a basis, we will now apply these techniques in a real-world situation focusing on the effect of ageing on cancer.

To our knowledge, we will now present the first symmetry based procedure for conducting model selection in a situation with actual experimental data. Using a time series consisting of the number of incidences of colon cancer for a wide range of ages [8], two initially inseparable candidate models corresponding to two different biological mechanisms are fitted to the data. By calculating their symmetries, we are able to distinguish between the two candidates and using this result we formulate a model selection criteria selecting the model which both fits the data *and* whose fit is invariant under the action of its symmetries on the data. Lastly, by transforming the data with the symmetries and then fitting the respective candidate models to the transformed data we are able to select one of the two candidate models based on the previously posed selection criteria.

Results

Two different models called the PLM and the IM-II describe the increase in incidences of colon cancer with age equally well

There are two plausible hypotheses for the increased risk of developing colon cancer at a high age. The first one is an accumulation of mutations due to ageing and the second one is a decline in the capacity of the immune system to repair damage with high age. These two biological mechanisms are the basis of the so called *power law model* (*PLM*) and the *immunological model* (*IM-II*) respectively [8]. Moreover, the fit of both these models agrees with the same time series data consisting of the number incidences of colon cancer for various ages really well (Fig 1) as $R^2_{\text{adj}} > 0.99$ in both cases (Tab 1). Thus, based on the fit alone it is not possible to distinguish between the models and thereby argue in favour one of these two biological mechanisms. To this end, the symmetries of these models are calculated in order to capture the unique properties of each class of curves which subsequently can be used to tell the two candidates apart.

Unique symmetries of the PLM and the IM-II makes it possible to distinguish between the models

To distinguish between the two candidates, the symmetries of each model have been calculated (Tab 2). The PLM has three so called infinitesimal generators of the Lie group denoted by $X_{1,0}$, $X_{1,1}$ and $X_{1,2}$ respectively (in fact these three span the Lie Algebra). The zeroth generator $X_{1,0}$ is *trivial* meaning that it maps points on a solution curve to the *same* solution curve. Both the other two generators are non-trivial meaning that they map points on a solution curve to *another* solution curve. Furthermore, both of the non-trivial generators of the PLM are uni-directional where $X_{1,1}$ transforms points in the t-direction while $X_{1,2}$ maps points in the R-direction. On the other hand, the IM-II has only one infinitesimal generator of the Lie group denoted by X_2 which is non-trivial. This generator is also a uni-directional operator mapping points in the t-direction

Table 1: *The fit of the two candidate models*. The columns from left to right are: the name of the model, the function describing the solution curve, the optimal fit R^2_{adj} and the optimal parameters. In the symmetry based methodology for model selection, it is the parameter A that is fitted to the transformed time series for both the PLM and the IM-II. The parameters γ in the former case and the parameter τ in the latter case are fixed to their optimal values displayed in this table. The parameter α in the IM-II is held constant meaning that it is not estimated and its value is set to $\alpha = 0.044$ [8]. The IM-II contains a "double exponential" for which we have chosen the notation "$\exp(e^x)$".

Model	*Curve*	*Optimal fit,* R^2_{adj}	*Optimal parameters*
Power law model (PLM)	$R(t) = At^{\gamma}$	0.9953	$\ln(A) \approx -14.0715$ $\gamma \approx 4.5280$
Immunological model (IM-II)	$R(t) = \dfrac{A}{\exp\left(e^{-\alpha(t-\tau)}\right) - 1}$	0.9921	$\ln(A) \approx 4.8731$ $\tau \approx 58.3782$ $\alpha = 0.044$ (Fixed)

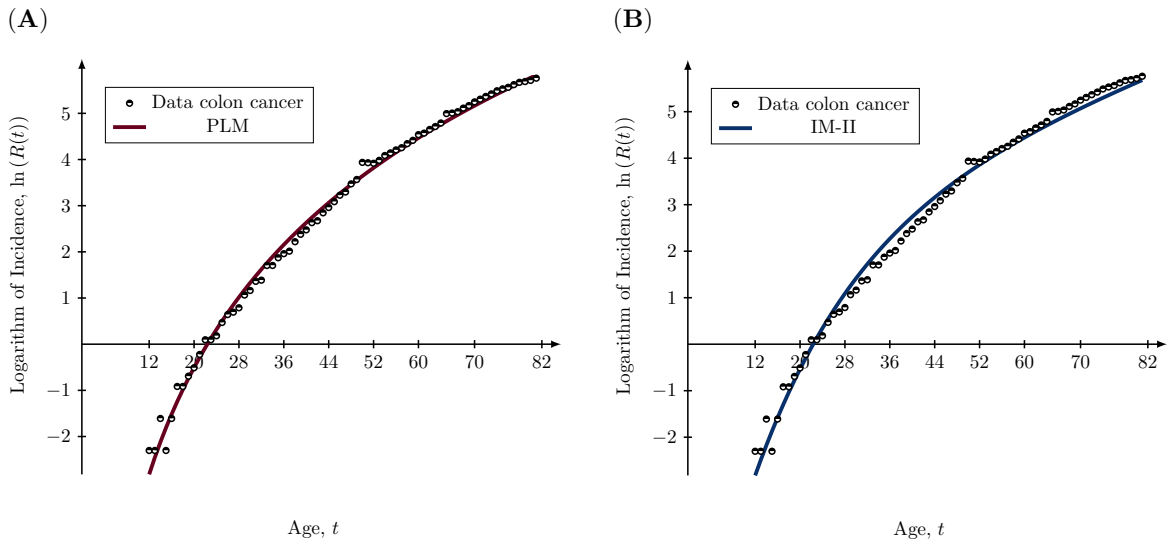

Figure 1: *Two different candidate models fit the data equally well*. Two models are fitted to the processed time series consisting of the logarithms of the incidences of colon cancer for ages in the range $(12, 83)$ years which is illustrated in two cases. (**A**) The PLM with optimal parameters $(\ln(A), \gamma) = (-14.0715, 4.5280)$ and fit $R^2_{\text{adj}} = 0.9953$. (**B**) The IM-II with optimal parameters $(\ln(A), \alpha, \tau) = (4.8731, 0.044, 58.3782)$ and fit $R^2_{\text{adj}} = 0.9921$.

(see section 1, 2, 4 and 5 in the supplementary material for more details on all symmetries). In fact, the usage of non-trivial as oppose to trivial symmetries in the model selection framework is a prerequisite and another necessary requirement is that *unique* symmetries of each candidate model are implemented as this enables the models to be distinguished. To make a direct comparison between the PLM and the IM-II, the non-trivial t-directional symmetries are implemented in the symmetry based framework for model selection.

The t-directional symmetries of the PLM and the IM-II makes it possible to distinguish between the two models. When the two optimal and geometrically similar curves obtained from the model fitting procedure of each candidate (Fig 1) are transformed with the unique t-directional symmetries of each model, the shape of the transformed solution curves are remarkable different in these two cases (Fig 2). Thus, the properties of the curves in each family are distinct and by transforming curves with the unique non-trivial symmetries of each model it is possible to differentiate between them. These symmetries are denoted by $\Gamma_{1,1}$ in case of the PLM and Γ_2 in case of the IM-II (Tab 2) as they are generated by the inifitesimal generators of the Lie group named by the same indices (i.e. $X_{1,1}$ and X_2 respectively). Subsequently, we will use these two symmetries as a basis for the model selection.

Table 2: *Summary of the symmetries of the two candidate models*. The models, their infinitesimal generators of the Lie group and the non-trivial symmetries are presented from left to right. The IM-II contains a "double exponential" for which we have chosen the notation "exp(e^x)".

Model	Infinitesimal Generators of the Lie group	Non-trivial Symmetries
PLM	$X_{1,0} = \left(\frac{t^{\gamma+1}}{\gamma}\right)\partial_t + Rt^{\gamma}\partial_R$ $X_{1,1} = t\partial_t$ $X_{1,2} = t^{\gamma}\partial_R$	$\Gamma_{1,1}(\epsilon) : (t, R) \mapsto (te^{\epsilon}, R)$ $\Gamma_{1,2}(\epsilon) : (t, R) \mapsto (t, R + \epsilon t^{\gamma})$
IM-II	$X_2 = e^{\alpha t}\exp\left(-e^{-\alpha(t-\tau)}\right)\partial_t$	$\Gamma_2(\epsilon) : (t, R) \mapsto \left(\tau - \frac{\ln\left(\ln\left(\left\lvert\alpha\epsilon - \exp\left(e^{-\alpha(t-\tau)}\right)\right\rvert\right)\right)}{\alpha}, R\right)$

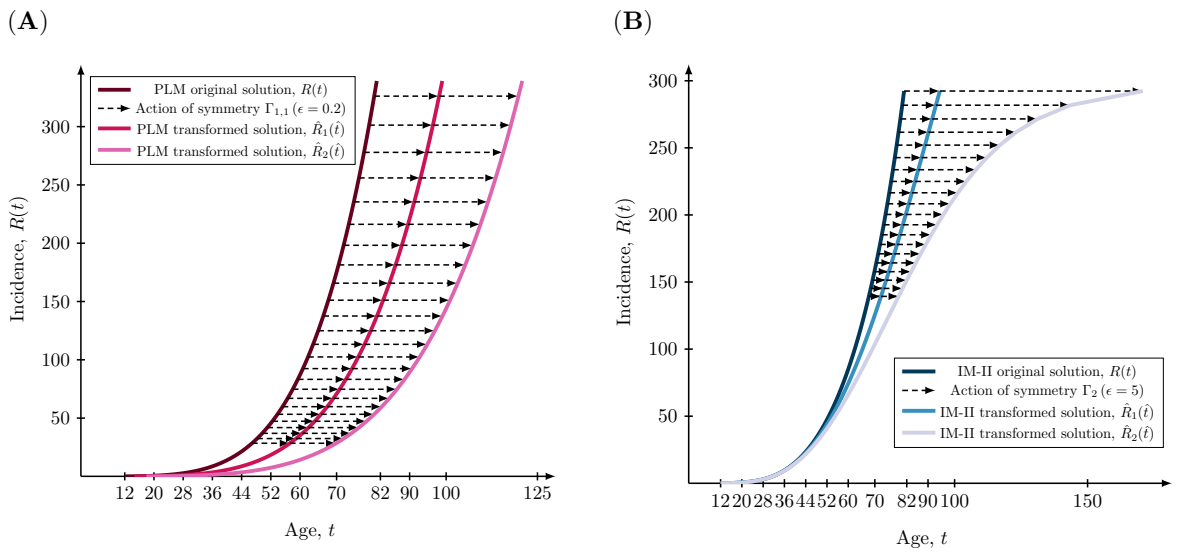

Figure 2: *Unique symmetries of the two candidate models can distinguish between them*. The action of the t-directional symmetries of the two candidate models are illustrated when the symmetries transform a solution curve twice with a fixed transformation parameter ϵ. The action of the symmetries is illustrated in two cases: (**A**) the scaling symmetry $\Gamma_{1,1}$ of the PLM with parameter $\epsilon = 0.2$ and (**B**) the symmetry Γ_2 of the IM-II with transformation parameter $\epsilon = 5$. The remaining two parameters of Γ_2 are fixed with the values $(\alpha, \tau) = (0.044, 58.3782)$.

The symmetries of the candidate models reveal that the PLM is a more realistic model than the IM-II

The key assumption for selecting a model based on its symmetries, is that be that the fit of the model should be invariant under symmetry transformation. Consequently, this criteria is added to the standard framework for statistical model selection which is based on the statement that the model that best fits the data is selected. This implies that the symmetry based framework selects one out of several candidate models by choosing the one that best fits the original data *and* that also can fit any new time series that is generated by transforming the original one using its symmetries. To quantify whether or not a the fit is invariant under the action of a given symmetry, we can simply plot the R^2_{adj}-value as a function of the transformation parameter ϵ which quantifies how much the original data is transformed by the symmetry at hand. To interpret the meaning of such a plot, let us investigate the expected outcomes of such an experiment.

The transformations of the data by a symmetry of a given model can either worsen the fit or leave it unchanged. In both scenarios, three objects are available to us namely a candidate model, its unique (with respect to the other candidates) symmetry and data in terms of a time series. Then, we proceed by transforming the data with the symmetry of the candidate model and thereafter the corresponding model is fitted to the transformed data which is quantified by the value of $R^2_{\text{adj}}(\epsilon)$. In the first case where the symmetry of the model is not manifest in the data, then the transformations of the symmetry will distort the data resulting in a new time series which the candidate model cannot describe. Also, the more the data is transformed the worse the fit will get which implies that the value of $R^2_{\text{adj}}(\epsilon)$ measuring the fit will decrease as a function of the transformation parameter ϵ. In the second case where the symmetry of the model is manifest in the data, then the fit of the model will be invariant under the action of the symmetry. This implies that any new time series that is generated by transforming the original one using the symmetry can be described by a particular solution curve from the family of solution curves of the candidate model. This implies that the value of $R^2_{\text{adj}}(\epsilon)$ will be constant as a function of the transformation parameter ϵ. Now, in the particular case of the two models of the increase in incidences of colon cancer during ageing it turns out that the symmetry based methodology favours one of the candidates over the other.

Based on the invariance of the fit to the transformed data, the PLM is more realistic than the IM-II. This is clear since when the original time series is transformed by the scaling symmetry $\Gamma_{1,1}$ another solution curve of the PLM can fit the transformed data really well (Fig 3A) while a corresponding curve of the IM-II fits the data transformed by the Γ_2-symmetry poorly in comparison (Fig 3B). The same observation can be quantified analysing the fit of the two models to the transformed data as functions of the transformation parameter. This analysis demonstrates that the $R^2_{\text{adj}}(\epsilon)$-value decreases with an increasing transformation parameter ϵ in the case of the IM-II while the same value is constant for the PLM (Fig 3C). According to the fundamental assumptions of the symmetry based methodology for model selection, the PLM is more realistic than the IM-II as it both fits the original time series *and* its fit is invariant under the action of its symmetry.

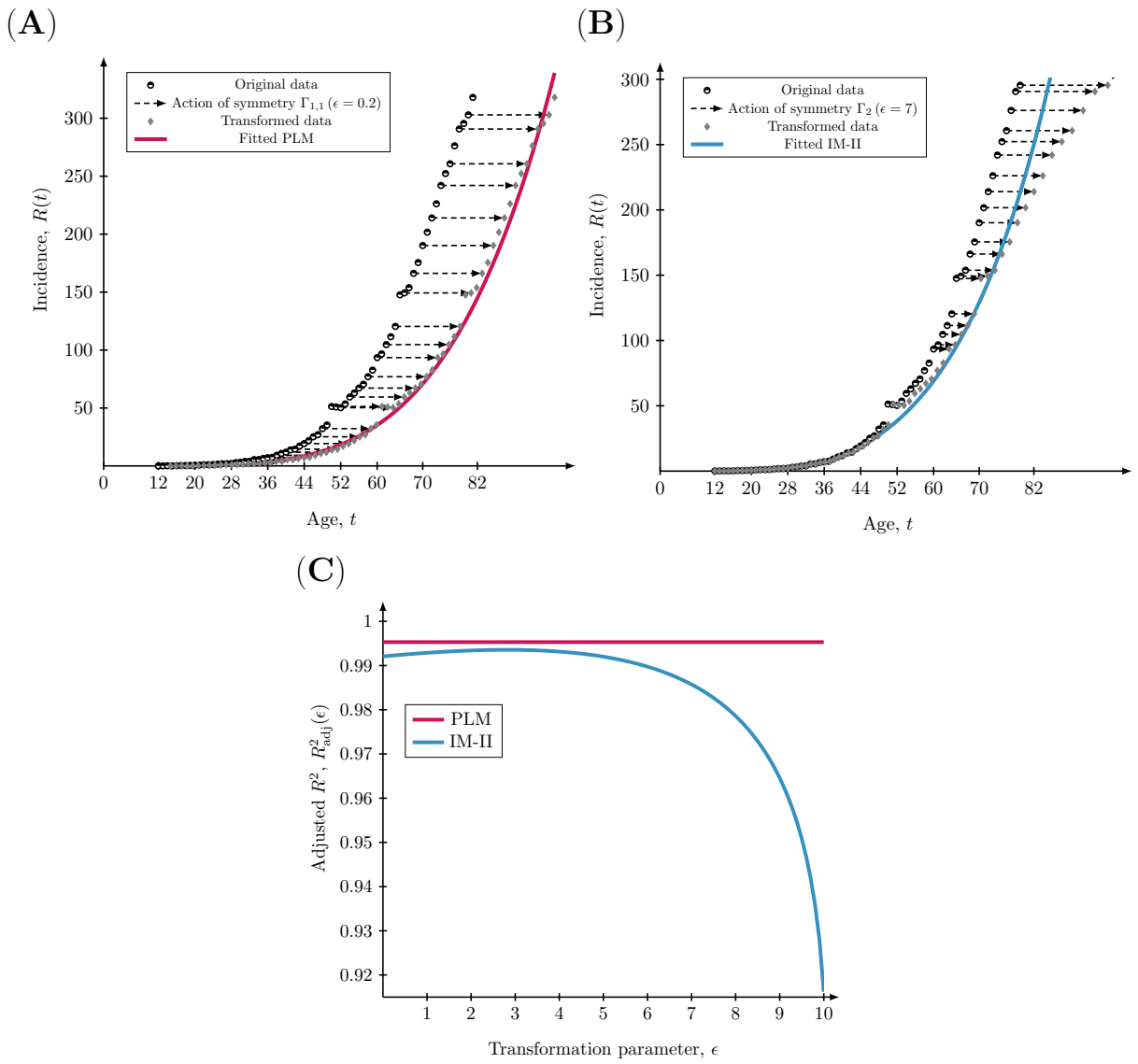

Figure 3: *The PLM is more realistic than the IM-II as its fit is invariant under transformations by its symmetry.* The two candidate models being the PLM and the IM-II is fitted to time series that are generated by transforming the original data using the t-directional symmetries of each model. This illustrated in two cases: (**A**) the fit of the PLM to a transformed time series generated by the scaling symmetry $\Gamma_{1,1}$ with transformation parameter $\epsilon = 0.2$ and (**B**) the fit of the IM-II to a transformed time series generated by the symmetry Γ_2 with transformation parameter $\epsilon = 7$ where the remaining two parameters are fixed to the values $(\alpha, \tau) = (0.044, 58.3782)$. (**C**) The fit of the PLM is invariant under the action of the scaling symmetry $\Gamma_{1,1}$ while the fit of the IM-II is *not* invariant under the action of its symmetry Γ_2. The value of $R^2_{\text{adj}}(\epsilon)$ remains constant in the case of the PLM fitted to multiple time series transformed by transformation parameters in the range $\epsilon \in [0, 10]$. In the case of the IM-II, the value of $R^2_{\text{adj}}(\epsilon)$ becomes more negative for increasing values of ϵ indicating that the fit is worsened the more the data is transformed. The same parameters that were used in the panels (A) and (B) are also used in panel (C).

Discussion

To our knowledge, this work is the first illustration of the use of symmetries in the selection of mechanistic models in biology based on *actual* experimental data. Here, we have demonstrated that symmetries can be used to select one among two different candidate models which both fit the same data with high accuracy. Initially, a time series of the increase in incidences of colon cancer due to ageing was presented together with two different but equally plausible biological explanations. These were the accumulation of harmful mutations during ageing encoded by the so called *power law model* (PLM) and a decreasing capacity of the immune system to repair damage with age described by the so called *immunological model* (IM-II). As these models could not be distinguished solely based on the fit as $R^2_{\text{adj}} > 0.99$ (Tab 1) in both cases, the symmetries of each candidate were derived in order to be able to differentiate between them. The two initially similar models could be distinguished by transforming their respective solution curves with their symmetries (Fig 2) and this result formed the basis for the model selection procedure. The fit of the IM-II was decreased when it was calibrated to time series obtained by transforming the original data using its symmetry while the fit of the PLM was invariant under transformations by its symmetry (Fig 3). By selecting the model that both fits the data and whose fit is invariant under the action of its symmetry, we concluded that the PLM is a more realistic description compared to the IM-II for explaining the increase in incidences of colon cancer during ageing.

There are three main considerations or limitations of the methodology which we now will discuss in chronological order. The first one concerns the philosophical assumption behind the idea of model selection, the second one is the difficulty of finding unique symmetries of the candidate models and the third is about the statistical assumption behind the fitting of the models to the data. Firstly, it should noted that in a standard model selection scenario a finite number of candidate models encoding different mechanisms are available and then the problem is to find the one candidate that best describes the available data. Although these models are often formulated based on the literature, it is often inevitable to make numerous assumptions in the construction of the models as there are "knowledge holes" in the study of any biological system. Due to the complexity of these systems, it is a possible that all proposed mechanisms or all candidate models in the selection phase if you will are incorrect. In such a case the symmetry based methodology will discard all models as the transformations of the data by the symmetries will distort the data in all cases provided that *unique symmetries of each candidate are known*.

Secondly, on this note, the difficulty of calculating unique symmetries poses a major obstacle in the usage of the methodology when modelling larger biological systems. On the one hand, the very task of finding the symmetries constitutes a bottleneck in biology as the symmetries consider all states (e.g. protein concentrations) and all variables (e.g. time and/or space) as variables in a high-dimensional geometrical space. Technically, to find the symmetries one must solve a high-dimensional PDE problem for which there are no general theoretical results and to do this efficiently an automated computer-based methodology must be developed. On the other hand, even if the symmetries of the candidate models are known it is also possible that some of the candidate models share *the same symmetries* which makes it impossible to distinguish between the candidate models using these particular symmetries. For instance, scaling symmetries such as $\Gamma_{1,1}$ of the PLM (Tab 2) are very common and another frequently occurring symmetry in biology is the so called *translation symmetry* given by $\Gamma_{3,1} : (t, R) \mapsto (t + \epsilon, R)$ for a single first order ODE. This type of symmetry is common to all *autonomous* differential equations which in the

context of ODEs correspond to equations where there is no dependence on time in the right hand side. An example of an autonomous model is the *IM-I* [8] corresponding to exponential growth mathematically formulated as $\mathrm{d}R/\mathrm{d}t = \alpha R$ (for details see section 1 in the supplementary material) and if two autonomous models are available in the model selection phase then it will not be possible to differentiate them from each other using the translation symmetry. Also, another problem is the effect of transformations by the symmetries on the properties of the original data which leads to a statistically motivated problem for the methodology.

Thirdly, transformations to data can distort the noise which ultimately affects the statistical method by which the model fitting is conducted. We have chosen the standard residual based maximum likelihood methodology in which the noise or error corresponding to the distance between the model and the data points is minimised. This method implicitly assumes an additive normally distributed noise and if the transformation of the noise by a symmetry changes its properties then it is statistically speaking not accurate to use the maximum likelihood approach when fitting a model to the transformed data. Hypothetically, this means that the action of a symmetry can distort the noise of the data resulting in a bad fit solely caused by the wrong choice of statistical method for calibrating the model to the transformed data. We acknowledge this statistical objection to which we do not present a solution as the focus in this work is on using symmetries in mechanistic modelling as oppose to develop novel statistical methods for model fitting.

In summary, this work showcases the power of symmetries in mechanistic modelling. Previously, the usefulness of this type of methodology for model selection had been demonstrated on synthetic data in the context of enzyme kinetics [7], but here the same type of argument illustrates that symmetries can play a vital role when actual data is available. Before the introduction of symmetries, the fundamental problem of selecting one out of numerous indistinguishable and potentially mutually exclusive models [1] was mainly solved based on the principle of Occam's razor where the smallest model (e.g. fewest states, parameters etc.) was chosen. However, if symmetries that encode physical or biological properties can be used to distinguish between seemingly indistinguishable models then this has vast implications for modelling as the underlying assumptions of models can be validated with higher accuracy. Better still, symmetries might be the means by which it is possible to insert meaning into the structure of the models which will allow us to avoid model selection completely or at least reformulate it as it currently stands. As the construction phase in modelling is limited by the knowledge of the system as well as the imagination of the modeller, an alternative methodology would be to use the symmetries to construct models as it is possible to formulate models starting from a single or multiple symmetries based on their invariants [4, 5]. In this way, it is possible to build in biological properties encoded by symmetries into the very structure of the model and in this way the construction phase would be based on physical principles rather than on unmotivated assumptions. Although the use of symmetries is in its infancy in biology, the same type of arguments have been used by Google's AI Deepmind to solve protein folding [9] where it was possible to re-construct the 3D structure of a protein merely based on its sequence of amino acids and this enormous success was in part enabled by assuming rotational invariance of the protein. Thus, the use of these methods has groundbreaking potential where it can be possible to build truly predictive and mechanistically validated models.

Materials and Methods

Data and fitting of the candidate models

The time series data has been collected from [Source needed][8]. It corresponds to the number of incidences of colon cancer of patients in the age span from zero to 83 years, and the original time series has 88 data points with three undefined NaN-values. Then, the data is processed in two steps. Firstly, all zero incidences are removed from the time series corresponding to all incidences below the ages of 12 years (i.e the first 12 data points are removed from the time series), as well as all undefined NaN-values which results in a time series with 73 data points. Secondly, in order to fit the models to the data, the logarithms of the remaining incidences are used.

To quantify the fit, the following version of the *adjusted* R^2 value is calculated

$$R^2_{\text{adj}} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right] \tag{1}$$

where

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(\ln(R_i) - \ln\left(\tilde{R}(t_i)\right)\right)^2}{\sum_{i=1}^{n}\left(\ln(R_i) - \overline{R}\right)^2} = 1 - \frac{\text{Variation between model and data}}{\text{Internal variation in data}}. \tag{2}$$

Note that in the definition of the R^2 value, it is the logarithm of the data R_i as well as the logarithm of the model $\tilde{R}(t_i)$ that is used. Also, the mean value $\overline{R}$ is calculated based on the *logarithm* of the incidences implying that

$$\overline{R} = \frac{\sum_{i=1}^{n}\ln(R_i)}{n}.$$

Moreover, the adjusted R^2 value penalises the number of parameters k, but for the two candidates the number of parameters is $k = 2$ in both cases. For both measures of the fit, i.e. R^2_{adj} and R^2, a perfect fit corresponds to the value 1 and the lower the value of R^2_{adj} the worse the fit (see section 3 of the supplementary material).

Accordingly, the logarithms of the two candidate models are fitted to the processed data. For the *power law model* (PLM) this means that the following model is calibrated to the processed data

$$\ln(R(t)) = \ln(A) + \gamma\ln(t). \tag{3}$$

For the *immunological model* (IM-II) the following model is fitted to the processed data

$$\ln(R(t)) = \ln(A) - \ln\left(\exp\left(e^{-\alpha(t-\tau)}\right) - 1\right). \tag{4}$$

It is important to emphasise that throughout the symmetry based model selection it is the parameters defining the family of each curve that are kept constant while the parameter defining an individual curve is varied. The parameter that defines the unique curve within each family of curves is the parameter A in both the case of the PLM (3) and the IM-II (4). The parameters that are kept fixed are γ in the case of the PLM and α and τ in the case of the IM-II. The values of these fixed parameters (Tab 1) are given by the optimal parameters obtained from the model calibration where the candidate models were fitted to the data.

Calculating the symmetries of each model

A symmetry is an operator which maps a solution curve to an *ordinary differential equation* (ODE) to another solution curve [5]. Let $\gamma = (t, R(t))$ be a solution curve to the ODE given by

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \omega(t, R)$$

where the function ω corresponds to the reaction term. Then a (point-wise) *symmetry* of this ODE is an operator of the type

$$\Gamma(\epsilon) : (t, R) \mapsto (\hat{t}(\epsilon), \hat{R}(\epsilon))$$

which maps a solution curve $\gamma = (t, R(t))$ to another solution curve $\hat{\gamma} = (\hat{t}, \hat{R}(\hat{t}))$. A restriction of this work is to focus on so called $\mathcal{C}^\infty$ symmetries parametrised by a single *transformation parameter* ϵ which implies that the target functions $\hat{t}(\epsilon)$ and $\hat{R}(\epsilon)$ are continuous functions of ϵ. Using this latter fact, it is possible to write the target functions as Taylor expansions locally around $\epsilon \approx 0$ as follows:

$$\begin{aligned}\hat{t}(\epsilon) &= t + \xi(t, R)\epsilon + \mathcal{O}\left(\epsilon^2\right), \\ \hat{R}(\epsilon) &= t + \eta(t, R)\epsilon + \mathcal{O}\left(\epsilon^2\right).\end{aligned}$$

The so called tangents ξ and η define the following vector field

$$X = \xi(t, R)\partial_t + \eta(t, R)\partial_R$$

which is referred to as the *infinitesimal generator of the Lie group* [5]. Using this local description of the action of the symmetry $\Gamma(\epsilon)$ it is possible to retrieve the global behaviour through the *exponential map* which is defined as follows

$$e^{\epsilon X} = \sum_{j=0}^{\infty} \frac{\epsilon^j}{j!} X^j.$$

More precisely, it is possible to generate a symmetry $\Gamma(\epsilon)$ by using its generator X according to the following equation

$$\Gamma(\epsilon) : (t, R) \mapsto \left(e^{\epsilon X} t, e^{\epsilon X} R\right).$$

Thus, it is sufficient to calculate the generator X since the corresponding symmetry $\Gamma(\epsilon)$ is obtained by the exponential map according to the above equation. The tangents ξ and η in the infinitesimal generator of the Lie group X are found by solving the so called *linearised symmetry condition* [5] defined as follows

$$\frac{\partial \eta}{\partial t} + \left(\frac{\partial \eta}{\partial R} - \frac{\partial \xi}{\partial t}\right)\omega(t, R) - \frac{\partial \xi}{\partial R}\omega(t, R)^2 = \xi\frac{\partial \omega}{\partial t} + \eta\frac{\partial \omega}{\partial R}.$$

A symmetry can be characterised as either *trivial* or *non-trivial* by using the *reduced characteristic* [5] denoted by $\overline{Q}$. For a first order ODE, it is defined as follows:

$$\overline{Q}(X) = \overline{Q}(t, R)\big|_{\xi,\eta \text{ defined by } X} = \eta(t, R) - \omega(t, R)\xi(t, R).$$

If $\overline{Q}(X) \equiv 0$ then the symmetry is *trivial* implying that it does not move any data points otherwise the symmetry is *non-trivial*. In the symmetry based methodology for model selection, only non-trivial symmetries are implemented.

Since the candidate models are formulated as functions or curves, their symmetries are found by firstly re-writing these functions as ODEs and secondly the linearised symmetry condition is solved in each case (for details see section 1 in the supplementary material).

A symmetry based framework for model selection

The symmetry based framework for selecting a candidate model is based on whether or not the fit of a candidate model is invariant under transformations of the data by its own symmetry. The implemented methodology is based on the previously developed symmetry based approach for model selection involving simulated data where the candidate models were based on the Hill equation in enzyme kinetics [7]. By transforming the original data with a particular symmetry $\Gamma(\epsilon)$ for a defined value of the transformation parameter ϵ, a new time series is obtained. If a curve from the same family of curves as the original one can be fitted to this new time series equally well as when the original curve was fitted to the original time series then the fit is said to be invariant under the action of the symmetry. In the case of the two candidate models, the parameter γ defines the family of solution curves for the PLM whereas the parameter τ defines the family of solution curves for the IM-II (note that the value of α is kept fixed and is therefore not considered to be a parameter but rather a constant). Thus, in both cases the original data is transformed with a transformation parameter ϵ by a symmetry $\Gamma(\epsilon)$ that is unique to each candidate model and then an optimal solution from these families of curves is determined by estimating the value of the parameter A. Then, the fit of this optimal curve to the transformed data is saved together with the corresponding transformation parameter ϵ. If the fit of a curve from a given candidate model is constant for all transformed time series determined by the parameter ϵ then the symmetry is manifest in the data and the fit is invariant. On the other hand, if the fit is decreased as the data is transformed by the symmetry, this implies that the symmetry is not manifest in the data and the effect of its transformations is that the time series is distorted. Lastly, the symmetry based methodology for model selection selects candidates which fit the data well and which fit to the (transformed) data is invariant under the action of its symmetries. It is worth emphasising that the fitting is done on the logarithm of the incidences, i.e. $\ln(R(t))$, while the transformations by the symmetries act on the incidences directly, i.e. on $R(t)$.

References

[1] P. Gerlee, "The model muddle: in search of tumor growth laws," *Cancer research*, vol. 73, no. 8, pp. 2407–2411, 2013.

[2] C. Hitchens, *God is not great: How religion poisons everything*. McClelland & Stewart, 2008.

[3] D. J. Gross, "The role of symmetry in fundamental physics," *Proceedings of the National Academy of Sciences*, vol. 93, no. 25, pp. 14256–14259, 1996.

[4] G. W. Bluman and S. Kumei, *Symmetries and differential equations*, vol. 81. Springer Science & Business Media, 1989.

[5] P. E. Hydon and P. E. Hydon, *Symmetry methods for differential equations: a beginner's guide*, vol. 22. Cambridge University Press, 2000.

[6] M. Golubitsky and I. Stewart, "Symmetry methods in mathematical biology," *São Paulo Journal of Mathematical Sciences*, vol. 9, no. 1, pp. 1–36, 2015.

[7] F. Ohlsson, J. Borgqvist, and M. Cvijovic, "Symmetry structures in dynamic models of biochemical systems," *Journal of the Royal Society Interface*, vol. 17, no. 168, p. 20200204, 2020.

[8] S. Palmer, L. Albergante, C. C. Blackburn, and T. J. Newman, "Thymic involution and rising disease incidence with age," *Proceedings of the National Academy of Sciences*, vol. 115, no. 8, pp. 1883–1888, 2018.

[9] E. Callaway, "'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures.," *Nature*, pp. 203–204, 2020.