

Equation learning to identify nano-engineered particle-cell interactions: an interpretable machine learning approach

Supplementary Information

S. T. Johnston¹ and M. Faria²

1. School of Mathematics and Statistics, The University of Melbourne, Victoria, Australia.

2. Department of Biomedical Engineering, The University of Melbourne, Victoria, Australia.

1 Equation learning convergence

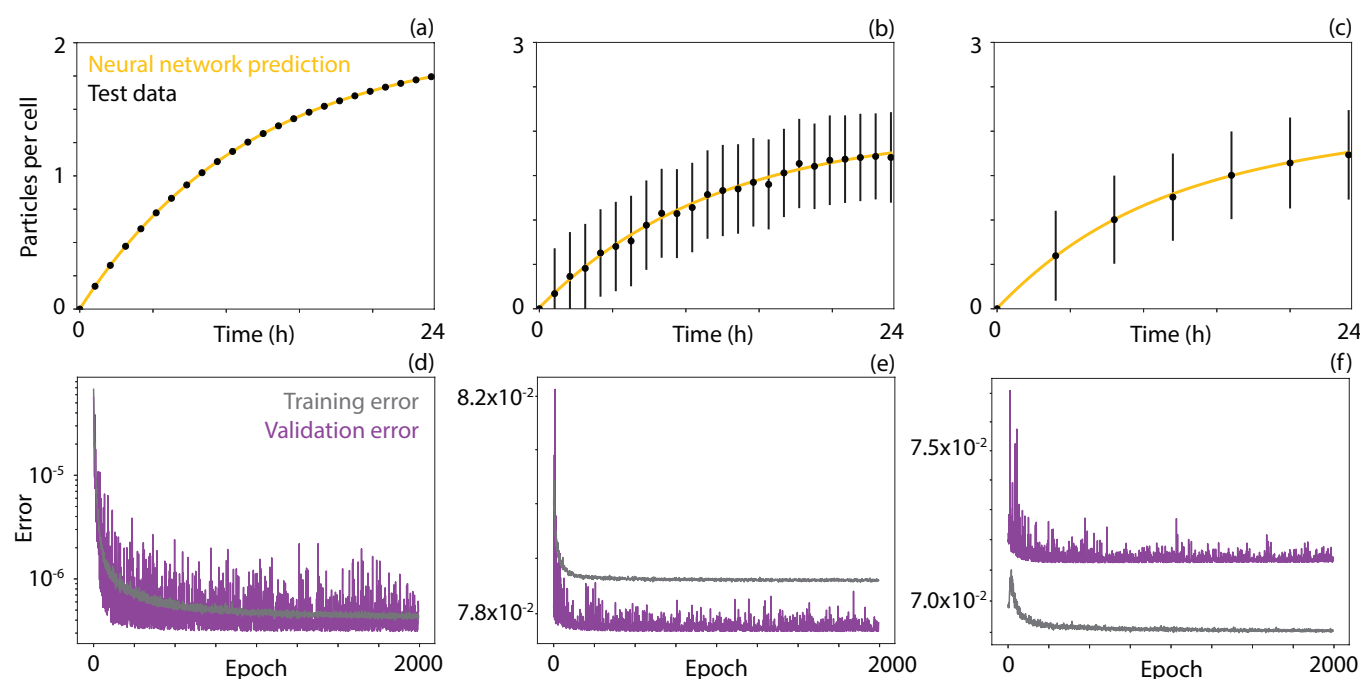


Figure 1: Equation learning framework applied to synthetic data with $\Omega_s(N, t) = d_0 - d_1 N$. (a)-(c) Comparison between the neural network prediction of the number of associated particles per cell (orange) and the test dataset (black). (d)-(f) Convergence of the neural network to the training dataset (grey) and the validation dataset (purple). The dataset contains 2,500 observations at each of (a),(b),(d),(e) 25 and (c),(f) 7 time points with Gaussian noise with a mean of zero and a standard deviation of (a),(d) 10^{-4} , (b),(c),(e),(f) 0.5.

We demonstrate that the neural network output matches the test dataset, and that the neural network converges to both the training and validation datasets in Figure 1 for the synthetic dataset generated with $\Omega_s(N, t) = d_0 - d_1 N$ for different numbers of observations and different levels of noise. The equation learning prediction of the test data and the rate of particle-cell association can be found in Figure 2 in the main document.

We again demonstrate that the neural network output matches the test dataset, and that the neural network converges to both the training and validation datasets in Figure 2 for the synthetic dataset generated with either $\Omega_s(N, t) = d_0 - d_1 N - d_2 N^2$ (Figures 2(a)-(b)) or $\Omega_s(N, t) = d_0 - d_1 N - d_2 N^2 - d_3 N^3$ (Figures

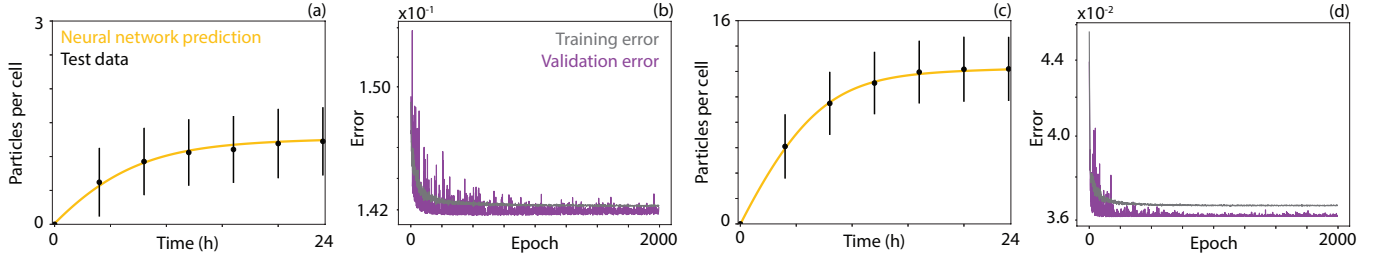


Figure 2: Equation learning framework applied to synthetic data with (a)-(b) $\Omega_s(N, t) = d_0 - d_1N - d_2N^2$ and (c)-(d) $\Omega_s(N, t) = d_0 - d_1N - d_2N^2 - d_3N^3$. (a),(c) Comparison between the neural network prediction of the number of associated particles per cell (orange) and the test dataset (black). (b),(d) Convergence of the neural network to the training dataset (grey) and the validation dataset (purple). The dataset contains 2,500 observations at each of 7 time points with Gaussian noise with a mean of zero and a standard deviation of (a)-(b) 0.5 and (c)-(d) 3.

2(c)-(d)). These results corresponds to the results presented in Figure 3 of the main document, where the equation learning prediction of the test data and association are shown.

2 Learned model parameters

As discussed in the main document, the learned model

$$\Omega_L(N, t) = d_0 - d_1N - d_2N^2 - \dots - d_mN^m,$$

can be recast into a form that contains key metrics of particle performance

$$\Omega_L(N, t) = \alpha C_0 S_A \left(1 - \delta_1 \left(\frac{N}{K} \right) - \delta_2 \left(\frac{N}{K} \right)^2 - \dots - \delta_m \left(\frac{N}{K} \right)^m \right),$$

where α is particle-cell affinity (m/s), C_0 is the initial particle concentration (m^{-3}), S_A is the surface area of the relevant cell type (m^2), K is the cell carrying capacity and δ_i is the relevant contribution of the i th component, subject to $\sum_i \delta_i = 1$. The δ_i and d_i terms are related via

$$d_i = \frac{\alpha C_0 S_A}{K^i} \delta_i.$$

The benefit of the recast form is that α and K have been previously identified as relevant metrics of particle performance, and the values of the δ_i allows for the relative contributions of the i th order saturation effect to be measured. We present the learned model parameters for the hyperbranched polymer particles in Table 1, and for the poly(methacrylic acid) particles in Table 2. We note that the cell carrying capacity value changes with the particle concentration for the hyperbranched polymer particles, which may suggest that the saturation in the number of associated particles per cell is due to a balance between particle-cell association and either particle recycling or particle degradation; however, we do not speculate on the precise biological mechanism here.

Experimental dataset	Learned model	Affinity (α)	Carrying capacity (K)	δ_1	δ_2	δ_3	Initial particle concentration (C_0)	Cell surface area (S_A)
PSMA(-) Control 1x	Model 2	2.05×10^{-12}	49.71	0.79	0.21	N/A	7.52×10^{18}	1.60×10^{-9}
PSMA(-) Control 4x	Model 2	6.57×10^{-13}	262.9	0.54	0.46	N/A	3.01×10^{19}	1.60×10^{-9}
PSMA(-) Control 16x	Model 3	5.18×10^{-13}	851.0	0.42	0.34	0.24	1.20×10^{20}	1.60×10^{-9}
PSMA(-) Peptide 1x	Model 3	1.63×10^{-12}	48.64	0.54	0.35	0.11	7.52×10^{18}	1.60×10^{-9}
PSMA(-) Peptide 4x	Model 2	5.36×10^{-13}	179.0	0.82	0.18	N/A	3.01×10^{19}	1.60×10^{-9}
PSMA(-) Peptide 16x	Model 3	3.76×10^{-13}	899.3	0.50	0.23	0.27	1.20×10^{20}	1.60×10^{-9}
PSMA(-) J591 1x	Model 2	2.69×10^{-12}	28.65	0.72	0.28	N/A	7.52×10^{18}	1.60×10^{-9}
PSMA(-) J591 4x	Model 2	7.31×10^{-13}	30.92	0.71	0.29	N/A	3.01×10^{19}	1.60×10^{-9}
PSMA(-) J591 16x	Model 1	1.23×10^{-13}	55.08	1.00	N/A	N/A	1.20×10^{20}	1.60×10^{-9}
PSMA(-) UG 1x	Model 2	1.85×10^{-12}	93.68	0.66	0.34	N/A	7.52×10^{18}	1.60×10^{-9}
PSMA(-) UG 4x	Model 3	8.27×10^{-13}	311.7	0.26	0.24	0.50	3.01×10^{19}	1.60×10^{-9}
PSMA(-) UG 16x	Model 3	7.64×10^{-13}	1441	0.39	0.26	0.35	1.20×10^{20}	1.60×10^{-9}
PSMA(+) Control 1x	Model 3	1.88×10^{-12}	56.80	0.60	0.29	0.11	7.52×10^{18}	1.60×10^{-9}
PSMA(+) Control 4x	Model 2	8.03×10^{-13}	332.2	0.37	0.63	N/A	3.01×10^{19}	1.60×10^{-9}
PSMA(+) Control 16x	Model 3	6.36×10^{-13}	1016	0.34	0.27	0.39	1.20×10^{20}	1.60×10^{-9}
PSMA(+) Peptide 1x	Model 3	8.19×10^{-12}	984.7	0.41	0.17	0.42	7.52×10^{18}	1.60×10^{-9}
PSMA(+) Peptide 4x	Model 2	4.17×10^{-12}	1523	0.38	0.62	N/A	3.01×10^{19}	1.60×10^{-9}
PSMA(+) Peptide 16x	Model 1	1.91×10^{-12}	1507	1.00	N/A	N/A	1.20×10^{20}	1.60×10^{-9}
PSMA(+) J591 1x	Model 4	4.29×10^{-12}	445.1	0.21	N/A	0.79	7.52×10^{18}	1.60×10^{-9}
PSMA(+) J591 4x	Model 2	4.74×10^{-12}	921.0	0.45	0.55	N/A	3.01×10^{19}	1.60×10^{-9}
PSMA(+) J591 16x	Model 2	3.36×10^{-12}	839.7	0.80	0.20	N/A	1.20×10^{20}	1.60×10^{-9}
PSMA(+) UG 1x	Model 2	2.34×10^{-12}	125.7	0.76	0.24	N/A	7.52×10^{18}	1.60×10^{-9}
PSMA(+) UG 4x	Model 3	9.78×10^{-13}	413.5	0.21	0.39	0.40	3.01×10^{19}	1.60×10^{-9}
PSMA(+) UG 16x	Model 3	9.69×10^{-13}	1926	0.54	0.28	0.18	1.20×10^{20}	1.60×10^{-9}

Table 1: Parameters for the learned models for the hyperbranched polymer particles. Note that there is an implicit scaling for K and α as the fluorescence signal for an individual particle is not known. Comparison of K and α values can only be done between datasets within this library, and not for external datasets.

Experimental dataset	Learned model	Affinity (α)	Carrying capacity (K)	δ_1	δ_2	δ_3	Initial particle concentration (C_0)	Cell surface area (S_A)
150 nm Coreshell	Model 2	1.12×10^{-7}	20.70	0.60	0.40	N/A	9.95×10^{12}	5.31×10^{-10}
214 nm Capsule	Model 2	2.25×10^{-8}	4.47	0.66	0.34	N/A	9.95×10^{12}	5.31×10^{-10}
282 nm Coreshell	Model 1	7.30×10^{-8}	22.85	1.00	N/A	N/A	9.95×10^{12}	5.31×10^{-10}
480 nm Capsule	Model 3	2.84×10^{-8}	12.76	0.60	0.31	0.09	9.95×10^{12}	5.31×10^{-10}
633 nm Coreshell	Model 1	9.06×10^{-8}	8.71	1.00	N/A	N/A	9.95×10^{12}	5.31×10^{-10}
1032 nm Capsule	Model 2	9.62×10^{-9}	2.77	0.66	0.34	N/A	9.95×10^{12}	5.31×10^{-10}

Table 2: Parameters for the learned models for the poly(methacrylic acid) particles.

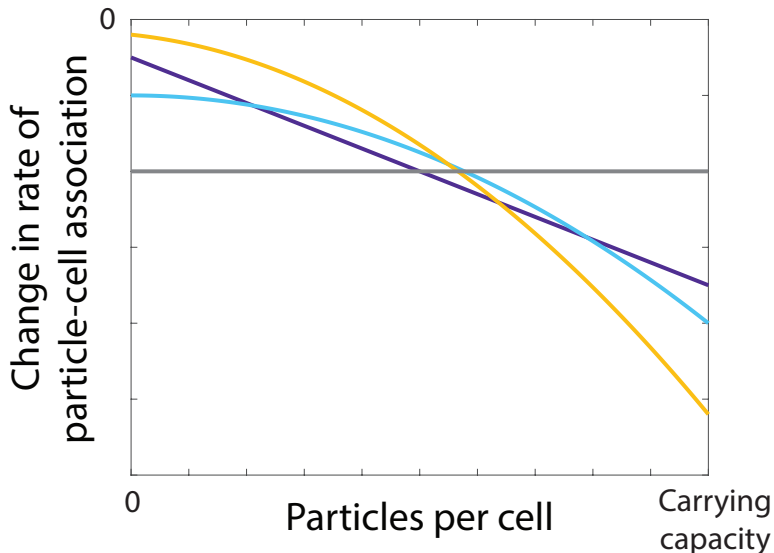


Figure 3: Illustrative examples of the change in the rate of particle-cell association as a function of the number of associated particles per cell for Model 1 (grey), Model 2 (purple), Model 3 (orange), and Model 4 (cyan). We observe that saturation effects are critical, as the rate of particle-cell association decreases with the number of particles per cell in each learned model.

3 Learned model transferability

A key question in machine learning is whether a learned model can be employed to describe data that the model was not trained on. To examine this, we take the most common learned model (Model 2) and determine whether the model can be used to describe datasets that have not been used previously in this study. These datasets include a comparison of the internalisation of carbon particles (fullerenes) at 10 μM and 25 μM concentrations by mammary carcinoma cells (MCF-7) [8]; a comparison of the cellular uptake of 20 nm amorphous silica particles for gastric cancer cells (MGC80-3) and cervical adenocarcinoma epithelial cells (HeLa) [10], and; a comparison of the association of metal phenolic network (MPN) particles with hyaluronic acid targeting ligands to human breast adenocarcinoma cells (MDA-MB-231) and human breast ductal carcinoma cells (BT-474) [7]. Full experimental details can be found in the previous literature for the carbon particles study [8], the amorphous silica particles study [10], and the hyaluronic acid MPN particles study [7].

For each dataset we fit Model 2 to the data via nonlinear least squares, and present a comparison between the experimental data and the model prediction in Figure 4. We observe that Model 2 is capable of describing each of the experimental datasets. This suggests that Model 2 is transferable and can be used beyond the datasets on which it was trained. Interestingly, for the datasets in Figures 4(d) and 4(e), Model 1 performs equally well as Model 2. This is consistent with our previous observation that Model 1 can be the most appropriate model to describe particle-cell association data.

4 Two-stage binding model

The processes by which nano-engineered particles are internalised are complicated, such that a single ordinary differential equation model of particle internalisation is unlikely to capture all of the relevant details. For example, numerous authors have presented models that represent the internalisation process as a two-stage process, where a particle first binds to the surface of a cell (via a receptor or a non-specific membrane

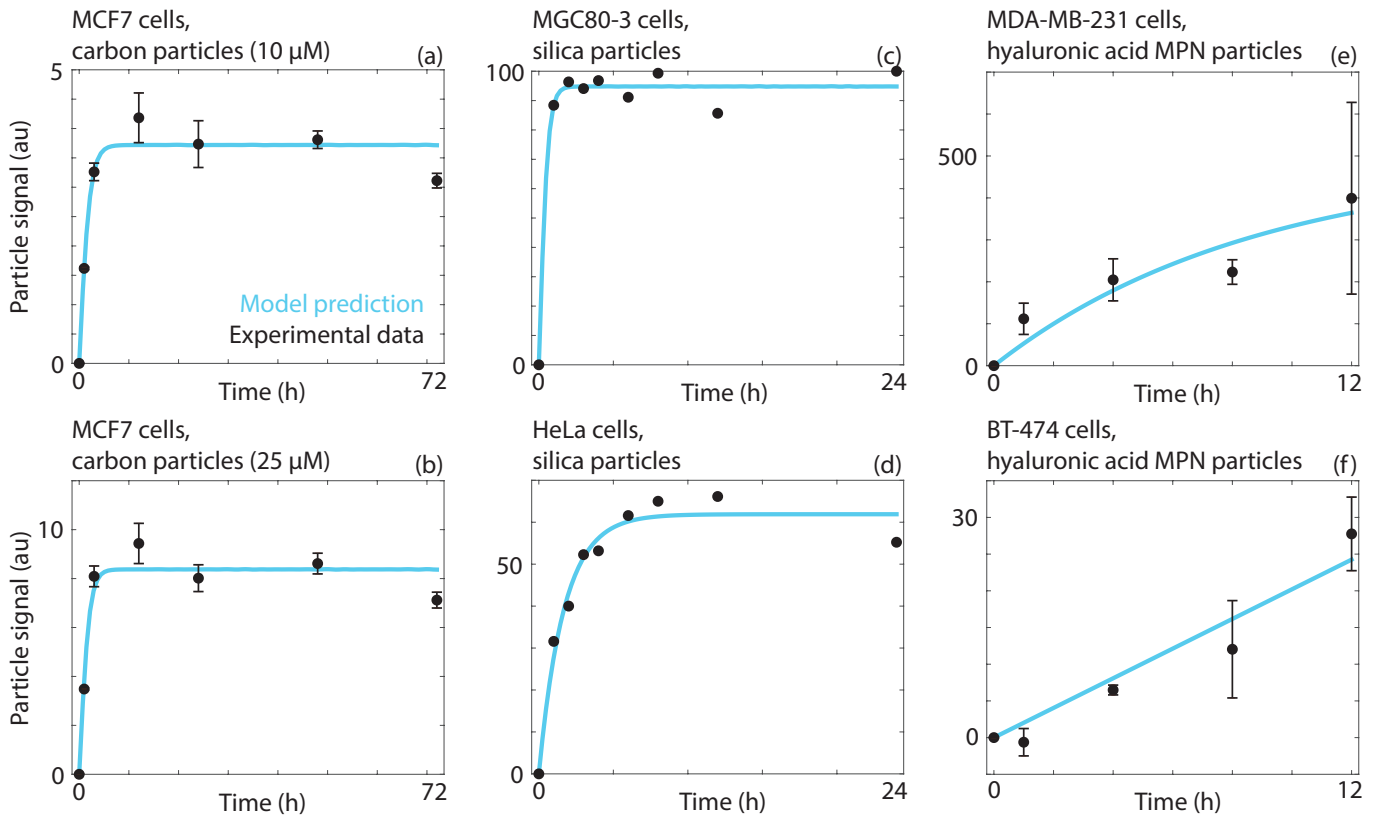


Figure 4: Transferability of the learned model (Model 2). Comparison between experimental data and predictions obtained from fitting Model 2 to the experimental data for (a) carbon particles at 10 μM concentration and MCF7 cells [8]; (b) carbon particles at 25 μM concentration and MCF7 cells [8]; (c) 20 nm amorphous silica particles and MGC80-3 cells [10]; (d) 20 nm amorphous silica particles and HeLa cells [10]; (e) hyaluronic acid MPN particles and MDA-MB-231 cells [7], and; (f) hyaluronic acid MPN particles and BT-474 cells [7]. Blue lines corresponds to model predictions and black dots correspond to the mean of the experimental data. Where available, error bars represent one standard deviation of the mean.

binding) and is subsequently internalised [1, 2, 3, 9]. These models can be represented mathematically as

$$\begin{aligned}\frac{dB(t)}{dt} &= \Omega^B(B(t), I(t)) - \Omega^I(B(t), I(t)), \\ \frac{dI(t)}{dt} &= \Omega^I(B(t), I(t)),\end{aligned}$$

where $B(t)$ is the average number of bound particles at time t , $I(t)$ is the average number of internalised particles at time t , $\Omega^B(B(t), I(t))$ is a function that describes the rate of particles binding to the cell and $\Omega^I(B(t), I(t))$ is a function that describes the rate of bound particles becoming internalised. Under appropriate choices of $\Omega^B(B(t), I(t))$ and $\Omega^I(B(t), I(t))$, it is straightforward to solve these equations to obtain predictions of $B(t)$ and $I(t)$. However, identifying $\Omega^B(B(t), I(t))$ and $\Omega^I(B(t), I(t))$ requires that bound and internalised particles can be distinguished. As various authors have discussed, this is not standard practice, and there are many issues that can be introduced via the choice of experimental techniques employed to distinguish between bound and internalised particles [5, 6].

Nonetheless, the difficulty of obtaining experimental data does not preclude examining whether the equation learning framework is capable of identifying the correct model. To achieve this, we can generate synthetic data from the above model, similar to the approach employed in Section 3.1 of the main manuscript. We can use this data to determine whether the equation learning framework can identify the forms of $\Omega^B(B(t), I(t))$ and $\Omega^I(B(t), I(t))$ when provided with data generated according to those choices of $\Omega^B(B(t), I(t))$ and $\Omega^I(B(t), I(t))$. We note that this approach will, in general, not work if the only data available is the number of associated particles (i.e. $N(t) = B(t) + I(t)$). This can be seen through

$$\frac{dN(t)}{dt} = \frac{dB(t)}{dt} + \frac{dI(t)}{dt} = \Omega^B(B(t), I(t)),$$

which is independent of the function representing internalisation, $\Omega^I(B(t), I(t))$. Moreover, $\Omega^B(B(t), I(t))$ may depend on both the number of bound and internalised particles, rather than the sum of the two, and this information is not present in the data.

We modify the equation learning framework such that it constructs $m = (n + 1)^2$ model components $\Omega = \{\omega_1(B_p, I_p, t), \dots, \omega_m(B_p, I_p, t)\}$. That is, the model components depend on the observations of both the number of bound and internalised particles. We choose model components that are polynomial terms of the form $B_p(t)^a I_p(t)^b$ for $0 \leq a \leq n$ and $0 \leq b \leq n$. We employ the following assumptions for the two neural networks:

- The numbers of both bound and internalised particles are non-negative.
- The rates of particle binding and particle internalisation are non-negative.

We do not include any assumptions about the presence of saturation effects, unlike for the single species model, as it is feasible that the rate of particle internalisation may increase with the number of bound particles. It is straightforward to constrain the neural network to produce such output. We employ a single hidden layer with a sigmoid activation function and non-negative biases and weights, alongside a linear output layer with a non-negative bias and weight [4].

Results

We generate data from the two-stage binding model by first selecting $\Omega_s^B(B(t), I(t)) = d_0 - d_1 B(t) - d_2 B^2(t)$ and $\Omega_s^I(B(t), I(t)) = d_3 B(t) + d_4 B^2(t)$. This represents an internalisation process where particle binding can saturate but particle internalisation does not saturate [1, 2, 3, 9]. As in Section 2.1 in the main manuscript, we generate synthetic data at time points corresponding to every hour over a total of 24 hours. We apply the modified equation learning framework to the synthetically-generated data and present the results in Figure 5. We first see that the learned model describes the test data accurately for both the bound and internalised particles. Crucially, we observe that the model learned from the data is of the correct form,

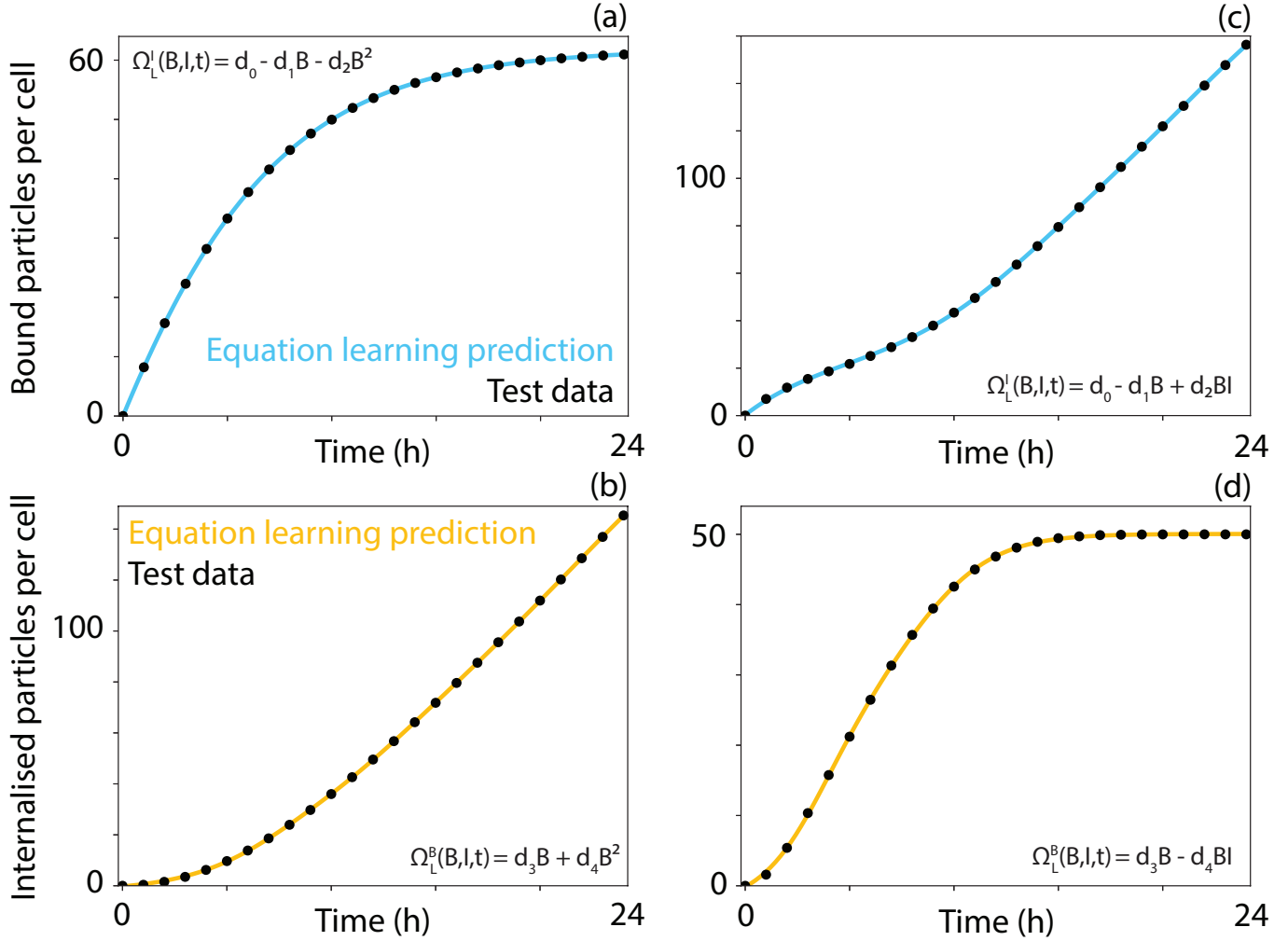


Figure 5: Equation learning framework applied to synthetic data with (a)-(b) $\Omega_s^B(B, I, t) = d_0 - d_1 B(t) - d_2 B^2(t)$ and $\Omega^I(B(t), I(t)) = d_3 B(t) + d_4 B^2(t)$ and (c)-(d) $\Omega_s^B(B, I, t) = d_0 - d_1 B(t) + d_2 B(t)I(t)$ and $\Omega^I(B(t), I(t)) = d_3 B(t) - d_4 B(t)I(t)$. Comparison between the equation learning prediction (cyan) and the test dataset (black) of the number of (a),(c) bound particles per cell and (b),(d) internalised particles per cell. The dataset contains 500 observations at each of 25 time points with Gaussian noise with a mean of zero and a standard deviation of 10^{-4} . Here we choose $n = 5$.

that is, $\Omega_L^B(B(t), I(t)) = d_0 - d_1B(t) - d_2B^2(t)$ and $\Omega_L^I(B(t), I(t)) = d_3B(t) + d_4B^2(t)$.

We next generate data from the two-stage binding model according to $\Omega_s^B(B(t), I(t)) = d_0 - d_1B(t) + d_2B(t)I(t)$ and $\Omega_s^I(B(t), I(t)) = d_3B(t) - d_4B(t)I(t)$. This represents an internalisation process where particle internalisation can saturate but particle binding does not saturate. We present the results obtained from the modified equation learning framework when applied to this synthetic data. Again, we observe that we obtain a close match between the test data and the learned model. Further, the learned model is of the form $\Omega_L^B(B(t), I(t)) = d_0 - d_1B(t) + d_2B(t)I(t)$ and $\Omega_L^I(B(t), I(t)) = d_3B(t) - d_4B(t)I(t)$, and hence matches the model used to generate the data. This indicates that the modified form of the equation learning framework is able to identify the correct model when data on both internalised and bound particles is available. Importantly, we also see that the equation learning framework can identify models when the constraints on the neural network are relaxed, as we observe that the first derivative of the number of bound particles is non-monotonic (i.e. it both increases and decreases) in Figure 5(c).

References

- [1] V. Belli, D. Guarnieri, M. Biondi, F. Della Sala, and P. A. Netti. Dynamics of nanoparticle diffusion and uptake in three-dimensional cell cultures. *Colloids Surf. B: Biointerfaces*, 149:7–15, 2017.
- [2] A. L. Doiron, B. Clark, and K. D. Rinker. Endothelial nanoparticle binding kinetics are matrix and size dependent. *Biotechnol. Bioeng.*, 108(12):2988–2998, 2011.
- [3] C. V. Dowling, P. M. Cevaal, M. Faria, and S. T. Johnston. On predicting heterogeneity in nanoparticle dosage. *bioRxiv*, 2022.
- [4] C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia. Incorporating functional knowledge in neural networks. *J. Mach. Learn. Res.*, 10(6), 2009.
- [5] L. I. FitzGerald and A. P. Johnston. It’s what’s on the inside that counts: Techniques for investigating the uptake and recycling of nanoparticles and proteins in cells. *J. Colloid Interface Sci.*, 587:64–78, 2021.
- [6] C. Gottstein, G. Wu, B. J. Wong, and J. A. Zasadzinski. Precise quantification of nanoparticle internalization. *ACS Nano*, 7(6):4933–4945, 2013.
- [7] Y. Ju, J. Cui, H. Sun, M. Müllner, Y. Dai, J. Guo, N. Bertleff-Zieschang, T. Suma, J. J. Richardson, and F. Caruso. Engineered metal-phenolic capsules show tunable targeted delivery to cancer cells. *Biomacromolecules*, 17(6):2268–2276, 2016.
- [8] M. Lucafo, S. Pacor, C. Fabbro, T. Da Ros, S. Zorzet, M. Prato, and G. Sava. Study of a potential drug delivery system based on carbon nanoparticles: effects of fullerene derivatives in mcf7 mammary carcinoma cells. *J. Nanopart. Res.*, 14(4):1–13, 2012.
- [9] C. Wilhelm, F. Gazeau, J. Roger, J. Pons, and J.-C. Bacri. Interaction of anionic superparamagnetic nanoparticles with cells: kinetic analyses of membrane adsorption and subsequent internalization. *Langmuir*, 18(21):8148–8155, 2002.
- [10] Y. Wu, W. Tang, P. Wang, C. Liu, Y. Yuan, and J. Qian. Cytotoxicity and cellular uptake of amorphous silica nanoparticles in human cancer cells. *Part. Part. Syst. Charact.*, 32(7):779–787, 2015.