Machine Learning Predictions of Drug Release from Polymeric Long Acting Injectables

Pauric Bannigan[1†], Florian Häse[2,3,4†], Matteo Aldeghi[2,3,4], Zeqing Bao[1], Alán Aspuru-Guzik[2,3,4,5*], Christine Allen[1*]

[1] Leslie Dan Faculty of Pharmacy, University of Toronto, Toronto, ON M5S 3M2, Canada.
[2] Chemical Physics Theory Group, Department of Chemistry, University of Toronto, Toronto, ON M5S 3H6, Canada.
[3] Department of Computer Science, University of Toronto, Toronto, ON M5S 3H6, Canada.
[4] Vector Institute for Artificial Intelligence, Toronto, ON M5S 1M1, Canada.
[5] Lebovic Fellow, Canadian Institute for Advanced Research, Toronto ON, M5S 1M1, Canada.

† P. Bannigan and F. Häse contributed equally to this work.

CORRESPONDING AUTHORS

*Christine Allen
Leslie Dan Faculty of Pharmacy,
University of Toronto,
Toronto, ON, M5S3M2,
Canada
Email: cj.allen@utoronto.ca

*Alan Aspuru-Guzik
Department of Chemistry,
University of Toronto,
Toronto, ON, M5S 3H6
Canada
E-mail: alan@aspuru.com

MAIN TEXT

Long acting injectables (LAI) are a class of advanced drug delivery systems that are designed to release their cargo over extended periods of time in order to achieve a prolonged therapeutic effect. LAIs that are amenable to parenteral administration can confer several advantages over conventional drug formulations, including increased patient compliance and bioavailability of drug [1]. Moreover, LAIs can be engineered to provide either local (e.g., Zilretta®) or systemic (e.g., Lupron Depot®) drug exposure over a prolonged period, making them ideal formulation strategies for the treatment of chronic diseases [2]. The unparalleled chemical and physical diversity afforded by polymers, makes polymer-based LAIs a particularly apt version of this drug delivery strategy. These systems can be engineered to entrap drug within a polymer matrix with release occurring via erosion, diffusion, or simultaneous erosion and diffusion mechanisms [3]. In addition to achieving sustained or controlled drug release, the encapsulation of drug into these polymeric matrices can often provide protection to the therapeutic cargo [4]. To date, various polymer-based LAI technologies administered via the intramuscular [5], subcutaneous [6], and intra-articular [7] routes have received regulatory approval (**Figure 1A**).

Despite the advantages associated with polymeric LAIs, their translation from bench to bedside remains non-trivial. In the past two decades, only about 30 polymeric LAI products have received regulatory approval, and this is in contrast to the thousands of conventional oral formulations approved in the same period [3,8]. Several pivotal challenges limit the clinical translation of polymeric LAIs. Firstly, there are few biodegradable polymer materials that are generally recognized as safe (GRAS) for parenteral administration. To date, the polymeric LAIs which have received clinical approval are largely based on a single polymer: poly(lactide-*co*-glycolide) (PLGA) [2,3,9]. The use of a material, such as PLGA, with an established safety profile may accelerate the regulatory approval of new LAI formulations. However, it is well recognized that polymer-drug compatibility significantly influences the performance of a formulation including drug loading capacity, drug release, and stability [10]. Given that each drug has its own unique physicochemical properties it is unlikely that any one polymer material is ideally suited for the formulation of all drugs. Thus, reliance on the relatively small subset of polymeric materials that have GRAS status likely restricts our ability to develop polymer-based LAIs for many classes of drugs. For drugs that display adequate compatibility to this limited number of GRAS polymer materials, there are a wide range of variables that must be optimized during the preparation of

LAIs. In addition, there are a number of distinct methods to prepare LAIs, and each method is associated with unique processing parameters that must be considered during formulation development [11]. Experimental evaluation of the number of variables that might be considered can soon become intractable. Moreover, changes to any of these variables has the potential to impact formulation performance, and the net effect of such alterations cannot be known *a priori.* During the development of LAIs emphasis is placed on achieving sustained or controlled drug release over extended periods of time. Often, initially promising formulations can fail at various stages during development due to unwanted drug release rates, making their re-formulation and re-evaluation necessary (**Figure 1B**). This trial-and-error based approach represents a significant bottleneck in LAI development [12].

To date, several strategies have been investigated to inform decision making and expedite the drug formulation development process. Mathematical models have been used to describe drug release and polymer degradation [13]. These models have greatly enhanced our understanding of the mechanisms underlying drug release and materials degradation; however, their application is generally limited to *post hoc* analysis. More recently, molecular dynamics simulations have been investigated [14]. These techniques have been useful in quantifying links between drug release rates and formulation parameters (including particle size and drug loading levels) [15]. While the development of these techniques is an active area of research, molecular scale simulations of entire drug delivery systems are currently computationally intractable [15]. Several studies have also investigated machine learning (ML) approaches [12]. ML techniques, in particular deep learning (DL), have recently been explored extensively for property prediction in chemistry and materials science [16–18]. Drug formulation development is associated with similar challenges to those currently being addressed using ML techniques in these areas. Regression models, such as those designed in this study, typically generate predictions by combining and transforming numerical input features through an array of stacked *layers* of *artificial neurons,* usually referred to as *neural networks* (NNs). The flexibility and established predictive power of NNs have led to the development of specific architectures for certain tasks, including convolutional NNs (i.e., image recognition), recurrent NNs (i.e., speech recognition), and graph neural networks (i.e., molecular property prediction).

The current study pursued the development and use of ML models for accurate prediction of drug release from polymeric LAIs (**Figure 1C**). Past efforts to predict *in vitro* drug release from LAIs

using ML have only examined models with a narrow scope of application. Szlęk et al. used NNs to predict drug release of proteins and peptides from PLGA-based MPs [19]. Small molecule studies have focused on MPs that include only a single drug formulation [20]. The limited scope of this previous work is not sufficient to evaluate whether ML might also allow us to design innovative polymeric LAIs. The current study was motivated by the ambitious goal of testing whether it is possible to develop a single ML model that can predict drug release profiles for arbitrary drug-polymer combinations. To evaluate this hypothesis, a dataset of 102 drug release profiles was curated from the literature and used as training data for various NN architectures. This training dataset included both MPs and cylinder-based implantable systems, 13 different small molecule drugs, and several commercially available polymers, including PLGA, polylactic acid (PLA), and polycaprolactone (PCL), of various molecular weights and repeat unit ratios. A second separate dataset containing 79 drug release profiles (made up of 33 new drug-polymer combinations) was also collected and used only for model evaluation. The NNs deployed in this study were trained to minimize root-mean-square deviation (RMSD) between their predictions and the experimental fractional drug release values. In this way, the models were developed to predict release iteratively per day. In addition, we also tested a model that would provide a single-shot estimate of the entire release curve by assuming a sigmoid functional form, NN_{sig}. In both cases, we tested NNs with one or two outputs. In the latter case, the final layer of the network output both the mean and variance of the predictions to provide an estimate of its own uncertainty, these models are referred to as *mean variance estimation (MVE) networks,* or MVE_{sig}, where the MVE network included parameterized sigmoid functions. Instead of predicting the experimental drug release values for a given input feature directly, MVEs parameterize a distribution from which the experimental drug release value is sampled at high likelihood (**Supplementary Equation S1**). More details on of each of these model architectures can be found in the SI.

The initial input features for the NNs were manually selected to account for both the physicochemical properties of individual formulations and release experiment conditions (i.e., amount of surfactant present in media). During the model development phase, NN structures with various numbers of input features (i.e., 9, 12, 13 and 25) were trained and evaluated using leave-one-group-out (LOGO) cross-validation. Briefly, the LOGO approach employed involved grouping the data into drug-polymer combinations, withholding one of these groups, training the model on the remaining groups and evaluating the accuracy of the predicted versus actual drug

release profiles for the withheld group in terms of the Pearson correlation coefficient (PCC) and RMSD (**Figure 1 C(ii); Supplementary Figure S1**). The relevance of the starting pool of 25 input features was determined through a series of analytical steps. For example, the magnitude of input weights was observed and served as a proxy for feature importance (**Figure 1 C(iii); Supplementary Figure S2**). Overall, the trained models predicted the release curves for different drug-polymer systems at accuracies that were in qualitative (PCC $\geq$ 0.95) and quantitative agreement (RMSD $\leq$ 0.2) with the experimentally obtained drug release profiles. In all cases the addition of parameterized sigmoid functions (i.e., NN_{sig} and MVE_{sig}) did not significantly improve the performance of models compared to the models without these constraints (**Figure 1C (ii)**).

Generally, the NN and MVEs predicted drug release with comparable overall performance, with RMSD values ranging from 0.1 – 0.2 (**Figure 1C(ii)**). Despite this comparable prediction accuracy, the MVEs yielded one potential advantage for future applications: prediction of fractional drug release along with an estimation of uncertainty in the prediction. The NN models also predict fractional drug release but afford no measure of prediction uncertainty. This prediction uncertainly feature could be useful in future applications as users can be more confident in predictions with low uncertainty. Moreover, model uncertainty can be used for active learning, where the ML model identifies the most informative data to collect in order to refine its predictions [21]. Overall, the 12 and 13 feature MVEs afforded the best prediction accuracy on the training set (**Figure 1C(ii)**). While the 12-feature MVE exhibited a marginal improvement versus the 13-feature NN (i.e., RMSD ~ 0.1 versus ~0.125), the 13-input feature configuration was chosen for the final model as it included an additional input feature that describes the physicochemical properties of the polymer **Figure 1C(iii)**.

To rigorously evaluate the generalizability of the selected 13-feature MVE, a pseudo-prospective study was conducted using a second dataset constructed from the published literature. The data from these published studies was digitized, tabulated, and used for prediction (**Table 1**). Most of the drug-polymer combinations in this external validation dataset were unseen by our model during training (i.e., 29/33). The 4/33 drug-polymer combinations that were present in the training set, differed considerably from previous entries in their input features, including drug loading capacity, surface to volume ratio of LAIs, and/or surfactant concentration in release media. The predicted release profiles for all formulations in the external validation dataset were compared with the experimental profiles (**Figure 2A**; **Supplementary Figure S3 and S4**). The overall prediction

accuracy of the model on the external validation dataset was quantified in terms of prediction accuracy and PCC (**Supplementary Equation S5**). In this way, the median prediction accuracy of the model was found to be 76%. This is a remarkable degree of accuracy for a ML model trained on such a small dataset (~ 100 release experiments), especially considering the diversity of drugs, polymers, drug-polymer combinations, and other LAI variables. Moreover, in many of the cases where the MVE prediction accuracy did not match the experimental release values, the uncertainty of the MVE was indicative of such lack of confidence, e.g., in the caffeine-PLGA, quercetin-PCL, triamcinolone acetonide-PLGA, and acetaminophen-PVL-*co*-PAVL predictions (**Figure 2A(i-iv)**). However, there were also a few exceptions where the predicted fractional drug release values were not only inaccurate, but where the error was also not captured by the estimated uncertainty: for example, letrozole-PLGA and lidocaine-PLGA (**Figure 2A(v-vi)**). In these cases, we postulate that the drug-polymer combinations may be beyond the predictive power of the current model design. This led us to a more thorough analysis of the predictions for the external validation set.

When the accuracy of the model on the external validation dataset was assessed by grouping predictions in terms of drug-polymer combination, it was found that the median prediction accuracy for most drug-polymer combinations was above the median prediction accuracy of the whole validation dataset. This observation had three exceptions: letrozole-PLGA, gefitinib-PLGA, and lidocaine-PLGA (**Figure 2B**). These three systems displayed a broad range of predictive accuracies, with lidocaine-PLGA having an especially large fraction of inaccurate predictions. The prediction accuracy for letrozole-PLGA ranged from ~0.95-0.55 (n=4), for gefitinib-PLGA from ~0.95 – 0.6 (n=5), and for lidocaine-PLGA from ~0.85 – 0.3 (n=20). Thus, there did not appear to be any drug-polymer group that the model was not able to predict well. Rather, there were several individual experiments for select drug-polymer groups that exhibited poor prediction accuracy. Similarly, there were some individual release profiles for other drug-polymer groups with a prediction accuracy that was significantly lower than the median value, i.e., paclitaxel-PVL-*co*-PAVL and triamcinolone acetonide-PVL-*co*-PAVL (**Figure 2B**).

When the prediction accuracy of the model was assessed in terms of drug (**Figure 2C(i)**) and polymer groups (**Figure 2C(ii)**), where polymers were grouped by the nature of the repeat unit (i.e., PLGA, PVA, etc.), a similar trend was observed. The median prediction accuracy of three drugs (letrozole, gefitinib, and lidocaine) were found to be below the overall median prediction accuracy of the model (**Figure 2C(i)**). These three drugs are all weak bases, and have the potential

to be at least partially ionized in physiologically relevant release media [22]. The ionization of these compounds could result in alterations in the physicochemical properties used as input features in the current model (i.e., LogP, TPSA, etc.). Moreover, these three drugs were also found to be the outliers for the abovementioned drug-polymer combination groupings, where each was prepared with PLGA (**Figure 2B**). This is noteworthy, as it has been proposed elsewhere that the hydrolysis of PLGA can result in a local decrease in pH [23], this phenomenon could have a profound effect on the release of such weakly basic drugs from the polymeric matrices. Moreover, changes in the ionization states of weakly basic and acidic compounds are well known to have a significant effect on drug solubility and dissolution rate [24,25]. Drug ionization during release experiments is a factor that warrants further investigation in future model iterations. Despite this observation, the prediction accuracy for several of the individual drug release profiles for letrozole, gefitinib, and lidocaine remained above the overall median accuracy of the model. Thus, indicating that perhaps values below the prediction median could be attributed to the polymers.

Hence, the effect of polymer group on prediction accuracy was also assessed (**Figure 2C(ii)**). Polymers were grouped by nature of repeat unit, molecular weight, and where applicable, lactic acid to glycolic acid ratio (i.e., for the PLGA systems). Interestingly, this grouping revealed that for almost half of the polymer groups the accuracy of the model was less than the overall median prediction accuracy (i.e., (PLGA, L/G= 65:35, Mw 57,700), (PLGA, L/G= 75:25, Mw 105,000), (PLGA, L/G= 50:50, Mw 44,000), (PVL-*co*-PAVL, Mw 150,000), (PLGA, L/G= 50:50, Mw 18,000), and (PLGA, L/G= 50:50, Mw 48,000)). Moreover, for two polymer groups the prediction accuracy for a significant number of the release experiments was less that 50% (i.e., (PLGA, L/G= 50:50, Mw 18,000), and (PLGA, L/G= 50:50, Mw 48,000)). Overall, the difference in the prediction accuracy for different polymers, and the same polymer with different attributes, suggests that the input features used in the final model do not fully capture subtle differences that can exist between polymers with the same repeat unit. Future additional studies are warranted to assess more optimal methods of numerically describing complex polymers, such as molecular fingerprints. Indeed polymer informatics is currently an active area of research, and methods to better capture the array of chemical and physical diversity of polymers are key to these research efforts [26]. Furthermore, it was also noted that the polymer groups with the lowest median accuracy (i.e., (PLGA, L/G= 50:50, Mw 18,000), and (PLGA, L/G= 50:50, Mw 48,000)) and the drug group with the lowest median accuracy (i.e., lidocaine) were extracted from the same original

publication. This observation highlights the challenge associated with training ML models with data extracted from published studies. While collecting all experimental data in-house provides maximum consistency, the volume of data accessible in such a way is limited. On the other hand, data collected from heterogenous literature sources may harbor hidden variability and inconsistencies that are detrimental to the development of accurate data-driven models. In fact, many laboratories have different experimental protocols for conducting release studies, and these subtle differences can result in distinct experimental outcomes. Better data transparency, and adaptation of automated experimental techniques could help to overcome this current hurdle in the application of ML tools in pharmaceutical formulation development as shown recently in several self-driving laboratory studies citations [27–30]. To this end, all trained models and datasets used in this study are openly available online (https://github.com/aspuru-guzik-group/to-be-released-upon-acceptance).

In summary, this study demonstrates that DL models can predict *in vitro* drug release from LAIs with a high degree of accuracy despite the minimal amount of training data (i.e., ~100 data samples). This is particularly promising for the application of ML in drug formulation development, and other areas of the pharmaceutical sciences, where large open access databases of experimental data are not readily available. Cutting-edge ML technologies are now freely available to pharmaceutical and material scientists, and the results obtained here demonstrate the potential for ML to expedite the development of innovative drug delivery technologies. Among the strengths of modern ML models are their ability to adapt to the prediction task at hand, and to provide uncertainty estimates reflecting model confidence. Although no model is entirely correct, here we show how useful models can be readily identified, even for prediction of complex properties. In addition, via model analysis, we have identified some key areas of improvement and likely sources of error. Future studies will focus on model optimization based on a larger training dataset with additional features, such as formulation processing features. This study, and its associated dataset, will foster the development of even more advanced, tailored, and accurate ML approaches for the prediction of drug release profiles. Overall, this proof-of-concept study sets the stage for a new era of data-driven development in advanced drug delivery research. The integration of these technologies into drug formulation development has the potential to not only inform decision making and accelerate development of promising drug formulations, but ultimately lead to the development of more innovative drug delivery platforms.

Figures/Tables:
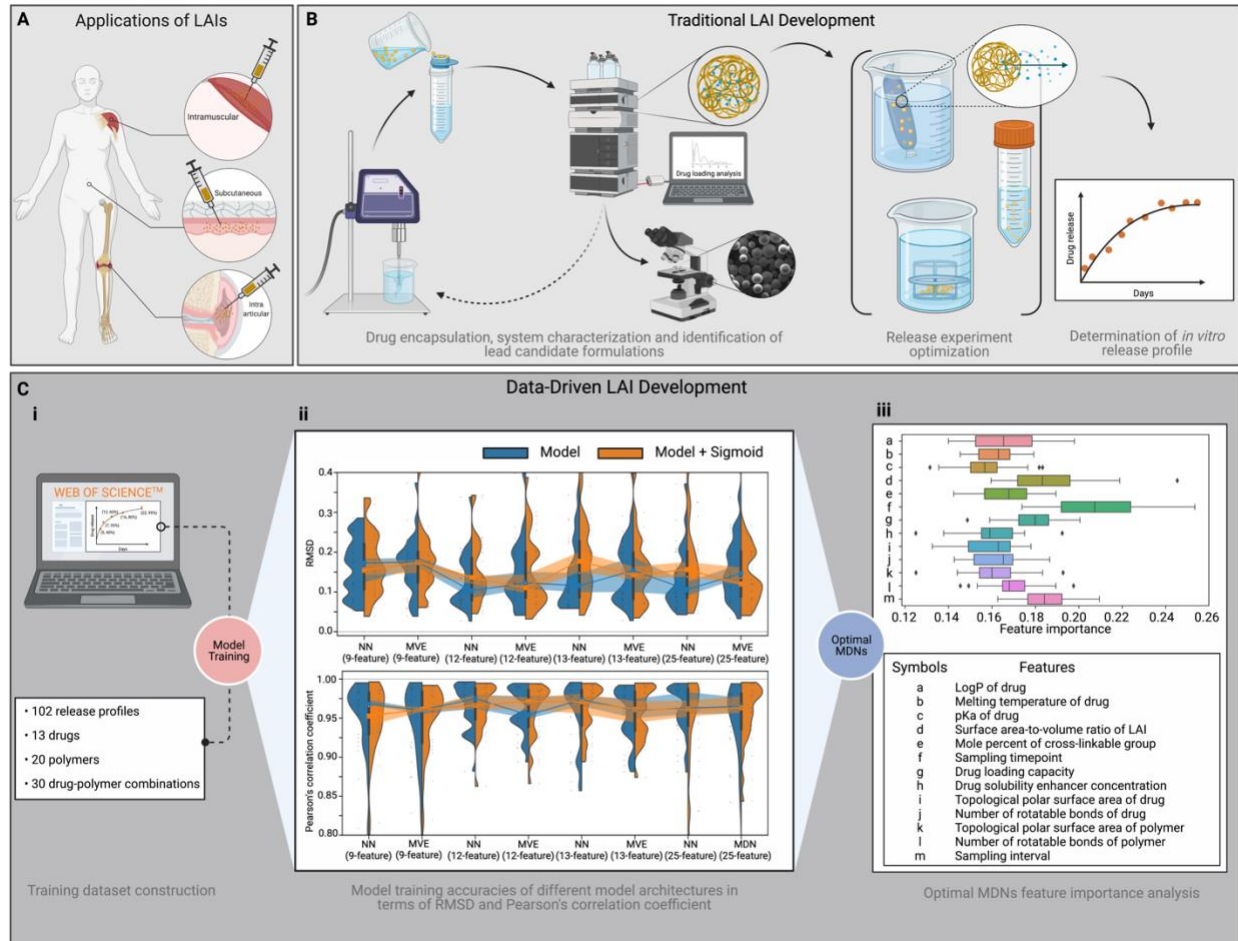
Figure 1: A) Scheme demonstrating selected routes of administration for FDA approved LAI formulations. B) Schematic depiction of a typical trial and error loop commonly observed during the development of LAIs, termed "Traditional LAI Development". C) Summary of the workflow employed here to train various ML models to predict drug release from LAIs, termed "Data-Driven LAI Development". C(i); Digitization of drug release curves obtained from the literature. C(ii); Summary of the overall results (i.e., RMSD and Pearson's correlation coefficient) obtained following training of various neural networks (NNs) and mean variance estimation (MVE) networks, without (blue) or with (orange) explicitly parameterized sigmoid functions, using the leave one group out cross validation method. C(iii): Summary of the final input features, and their relative importance, following retrospective analysis of the optimal MVE model.

Table 1. *Summary of drugs, polymers, and LAI properties for systems included in the external validation set.*

Drug	Polymer (average molecular weight)	Drug loading capacity	Particle LAI size (average)	Spherical or cylindrical LAI
Dexamethasone (DEX)	PEA (49 kD) PLGA (28 kD; L/G = 50/50) PLGA (13 kD; L/G = 50/50)	15% 8% 8%	14 µm 20 µm 20 µm	Spherical
Gefitinib (GEF)	PLGA (44 kD; L/G = 50/50)	3 - 8%	5 - 130 µm	Spherical
Cannabidiol (CBD)	PLGA (12 kD; L/G = 50/50) PLGA (51 kD; L/G = 50/50) PCL (42.5 kD)	8% 5% 9%	24 µm 24 µm 50 µm	Spherical
Tetrahydrocannabinol (THC)	PCL (42.5 kD)	8%	50 µm	Spherical
Caffeine (CAF)	PLGA (46 kD; L/G = 50/50)	6%	62 - 94 µm	Spherical
Lidocaine (LDC)	PLGA (12 kD; L/G = 50/50) PLGA (48 kD; L/G = 50/50) PLGA (10 kD; L/G = 70/30)	1 - 18% 1 - 2% 5 - 42%	3 - 9 µm 4 µm 5 - 100 µm	Spherical
Etoricoxib (ETC)	PCL (10 kD)	3%	16 µm	Spherical
Quercetin (QRC)	PCL (14 kD)	3 - 4%	61 – 100 µm	Spherical
Paclitaxel (PTX)	PLA (106 kD) PLGA (57.5 kD; L/G = 50/50) PLGA (10.5 kD; L/G = 75/25) PVL-*co*-PAVL (15 kD) PVL-*co*-PAVL (32 kD) PVL-*co*-PAVL (39 kD)	2% 1 - 2% 1 - 2% 10 % 10 - 20% 10 - 20%	0.6 µm 0.4 - 0.6 µm 0.3 - 0.7 µm 3 mm×2 mm (D×L) 3 mm×2 mm (D×L) 3 mm×2 mm (D×L)	Spherical Spherical Spherical Cylinder Cylinder Cylinder
Triamcinolone acetonide (TAA)	PLGA (18 kD; L/G = 50/50) PLGA (54 kD; L/G = 50/50) PVL-*co*-PAVL (15 kD) PVL-*co*-PAVL (32 kD) PVL-*co*-PAVL (39 kD)	5% 5% 10% 10 - 20% 10%	73 µm 71 µm 3 mm×2 mm (D×L) 3 mm×2 mm (D×L) 3 mm×2 mm (D×L)	Spherical Spherical Cylinder Cylinder Cylinder
Triamcinolone hexacetonide (TAH)	PVL-*co*-PAVL (15 kD) PVL-*co*-PAVL (39 kD)	10% 10%	3 mm×2 mm (D×L)	Cylinder
Curcumin (CCM)	PVL-*co*-PAVL (15 kD)	10%	3 mm×2 mm (D×L)	Cylinder
Acetaminophen (ACE)	PVL-*co*-PAVL (15 kD) PVL-*co*-PAVL (39 kD)	10%	3 mm×2 mm (D×L)	Cylinder
Letrozole (LTZ)	PLGA (57.7 kD; L/G = 65/35)	12 - 20%	0.6 - 1 µm	Spherical
Diprophylline (DPP)	PLGA (46 kD; L/G = 50/50)	5 - 7%	63 - 296 µm	Spherical
Ibuprofen (IBP)	PLGA (46 kD; L/G = 50/50)	14 - 20%	91 - 383 µm	Spherical
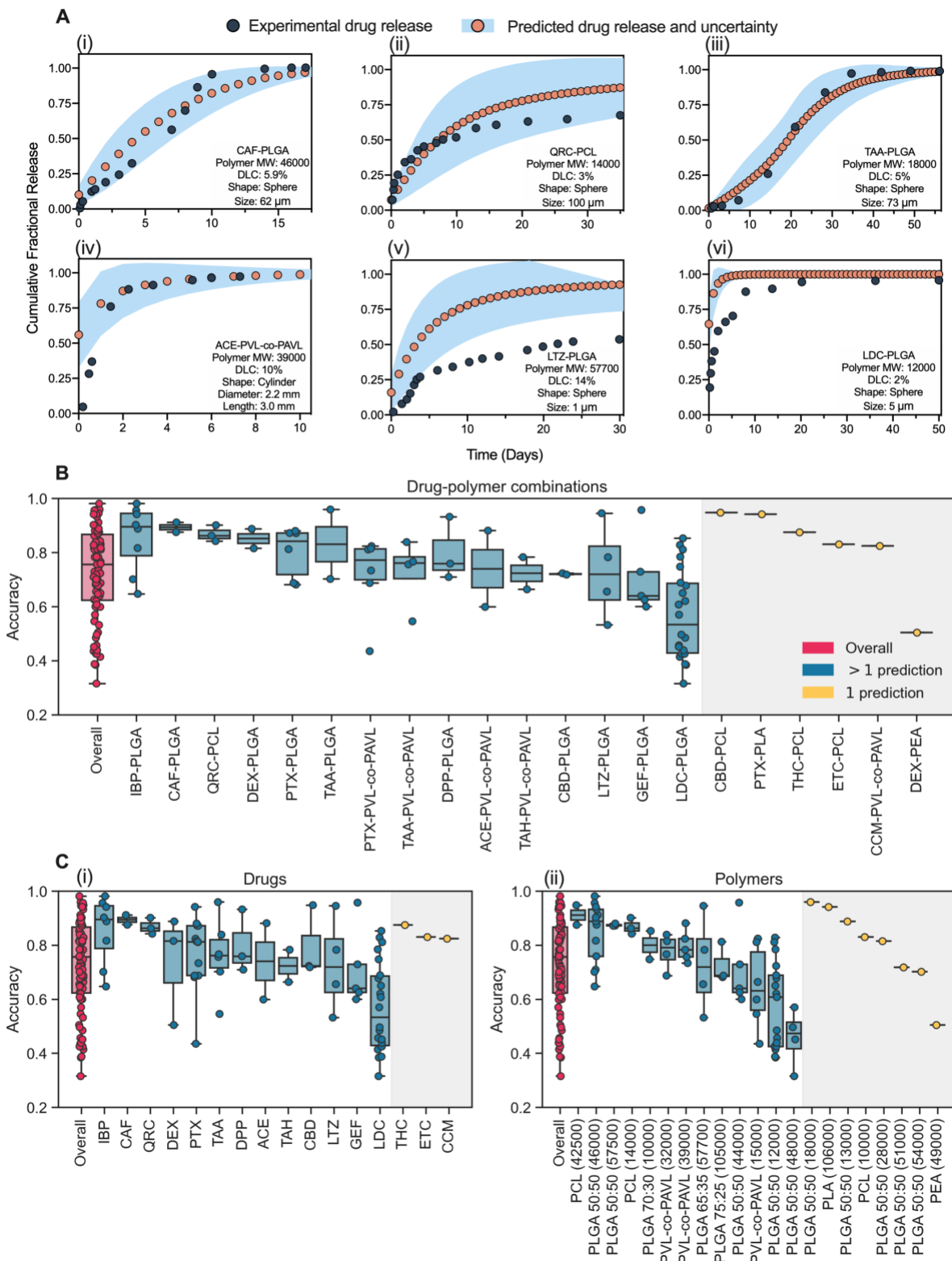
Figure 2. *A) Examples of experimental release profiles in comparison to the predicted profiles and associated uncertainty. B) Summary of the overall accuracy of the model in comparison to the accuracy for the individual drug-polymer combinations, C(i) drugs, and C(ii) polymers. The accuracies based on more than one prediction are ranked, in the left panel, in terms of their median values and the accuracies determined on only one prediction are ranked on the right.*

REFERENCES:

[1] N.C. Brigham, R.-R. Ji, M.L. Becker, Degradable polymeric vehicles for postoperative pain management, Nature Communications. 12 (2021) 1367. https://doi.org/10.1038/s41467-021-21438-3.

[2] J. Ghitman, E.I. Biru, R. Stan, H. Iovu, Review of hybrid PLGA nanoparticles: Future of smart drug delivery and theranostics medicine, Materials & Design. 193 (2020) 108805. https://doi.org/10.1016/j.matdes.2020.108805.

[3] M.N. O'Brien, W. Jiang, Y. Wang, D.M. Loffredo, Challenges and opportunities in the development of complex generic long-acting injectable drug products, Journal of Controlled Release. 336 (2021) 144–158. https://doi.org/10.1016/j.jconrel.2021.06.017.

[4] J. Siepmann, F. Siepmann, Microparticles Used as Drug Delivery Systems, in: W. Richtering (Ed.), Smart Colloidal Materials, Springer, Berlin, Heidelberg, 2006: pp. 15–21. https://doi.org/10.1007/3-540-32702-9_3.

[5] S.M. Agnihotri, P.R. Vavia, Pharmacokinetics of intramuscular microparticle depot of valdecoxib in an experimental model, Drug Dev Ind Pharm. 35 (2009) 1043–1047. https://doi.org/10.1080/03639040902762979.

[6] D.B. Shenoy, R.J. D'Souza, S.B. Tiwari, N. Udupa, Potential applications of polymeric microsphere suspension as subcutaneous depot for insulin, Drug Dev Ind Pharm. 29 (2003) 555–563. https://doi.org/10.1081/ddc-120018644.

[7] V.B. Kraus, P.G. Conaghan, H.A. Aazami, P. Mehra, A.J. Kivitz, J. Lufkin, J. Hauben, J.R. Johnson, N. Bodick, Synovial and systemic pharmacokinetics (PK) of triamcinolone acetonide (TA) following intra-articular (IA) injection of an extended-release microsphere-based formulation (FX006) or standard crystalline suspension in patients with knee osteoarthritis (OA), Osteoarthritis and Cartilage. 26 (2018) 34–42. https://doi.org/10.1016/j.joca.2017.10.003.

[8] K. Park, S. Skidmore, J. Hadar, J. Garner, H. Park, A. Otte, B.K. Soh, G. Yoon, D. Yu, Y. Yun, B.K. Lee, X. Jiang, Y. Wang, Injectable, long-acting PLGA formulations: Analyzing PLGA and understanding microparticle formation, Journal of Controlled Release. 304 (2019) 125–134. https://doi.org/10.1016/j.jconrel.2019.05.003.

[9] D.J. Hines, D.L. Kaplan, Poly (lactic-co-glycolic acid) controlled release systems: experimental and modeling insights, Crit Rev Ther Drug Carrier Syst. 30 (2013) 257–276.

[10] Z. Bao, S. Jung, J. Bufton, J.C. Evans, D.J. Aguiar, C. Allen, Poly(δ-valerolactone-co-allyl-δ-valerolactone) cross-linked microparticles: Formulation, characterization and biocompatibility, Journal of Pharmaceutical Sciences. 110 (2021) 2771–2777. https://doi.org/10.1016/j.xphs.2021.03.009.

[11] D. Cun, C. Zhang, H. Bera, M. Yang, Particle engineering principles and technologies for pharmaceutical biologics, Advanced Drug Delivery Reviews. 174 (2021) 140–167. https://doi.org/10.1016/j.addr.2021.04.006.

[12] P. Bannigan, M. Aldeghi, Z. Bao, F. Häse, A. Aspuru-Guzik, C. Allen, Machine learning directed drug formulation development, Advanced Drug Delivery Reviews. (2021). https://doi.org/10.1016/j.addr.2021.05.016.

[13] I. Irurzun-Arana, C. Rackauckas, T.O. McDonald, I.F. Trocóniz, Beyond Deterministic Models in Drug Discovery and Development, Trends in Pharmacological Sciences. 41 (2020) 882–895. https://doi.org/10.1016/j.tips.2020.09.005.

[14] O.M.H. Salo-Ahen, I. Alanko, R. Bhadane, A.M.J.J. Bonvin, R.V. Honorato, S. Hossain, A.H. Juffer, A. Kabedev, M. Lahtela-Kakkonen, A.S. Larsen, E. Lescrinier, P. Marimuthu,

M.U. Mirza, G. Mustafa, A. Nunes-Alves, T. Pantsar, A. Saadabadi, K. Singaravelu, M. Vanmeert, Molecular Dynamics Simulations in Drug Discovery and Pharmaceutical Development, Processes. 9 (2021) 71. https://doi.org/10.3390/pr9010071.

[15] T. Casalini, Not only in silico drug discovery: Molecular modeling towards in silico drug delivery formulations, Journal of Controlled Release. 332 (2021) 390–417. https://doi.org/10.1016/j.jconrel.2021.03.005.

[16] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, S. Zhao, Applications of machine learning in drug discovery and development, Nat Rev Drug Discov. 18 (2019) 463–477. https://doi.org/10.1038/s41573-019-0024-5.

[17] R. Pollice, G. dos Passos Gomes, M. Aldeghi, R.J. Hickman, M. Krenn, C. Lavigne, M. Lindner-D'Addario, A. Nigam, C.T. Ser, Z. Yao, A. Aspuru-Guzik, Data-Driven Strategies for Accelerated Materials Design, Acc. Chem. Res. 54 (2021) 849–860. https://doi.org/10.1021/acs.accounts.0c00785.

[18] F. Häse, L.M. Roch, P. Friederich, A. Aspuru-Guzik, Designing and understanding light-harvesting devices with machine learning, Nat Commun. 11 (2020) 4587. https://doi.org/10.1038/s41467-020-17995-8.

[19] J. Szlęk, A. Pacławski, R. Lau, R. Jachowicz, A. Mendyk, Heuristic modeling of macromolecule release from PLGA microspheres, Int J Nanomedicine. 8 (2013) 4601–4611. https://doi.org/10.2147/IJN.S53364.

[20] Y. Li, A.M. Rauth, X.Y. Wu, Prediction of kinetics of doxorubicin release from sulfopropyl dextran ion-exchange microspheres using artificial neural networks, European Journal of Pharmaceutical Sciences. 24 (2005) 401–410. https://doi.org/10.1016/j.ejps.2004.12.005.

[21] D. Reker, G. Schneider, Active-learning strategies in computer-assisted drug discovery, Drug Discovery Today. 20 (2015) 458–465. https://doi.org/10.1016/j.drudis.2014.12.004.

[22] M. Duvvuri, Y. Gong, D. Chatterji, J.P. Krise, Weak Base Permeability Characteristics Influence the Intracellular Sequestration Site in the Multidrug-resistant Human Leukemic Cell Line HL-60*, Journal of Biological Chemistry. 279 (2004) 32367–32372. https://doi.org/10.1074/jbc.M400735200.

[23] H. Liu, E.B. Slamovich, T.J. Webster, Less harmful acidic degradation of poly(lacticco-glycolic acid) bone tissue engineering scaffolds through titania nanoparticle addition, Int J Nanomedicine. 1 (2006) 541–545. https://doi.org/10.2147/nano.2006.1.4.541.

[24] U. Domańska, A. Pobudkowska, A. Pelczarska, Ł. Żukowski, Modelling, solubility and pKa of five sparingly soluble drugs, International Journal of Pharmaceutics. 403 (2011) 115–122. https://doi.org/10.1016/j.ijpharm.2010.10.034.

[25] T.R. Bates, J.M. Young, C.M. Wu, H.A. Rosenberg, pH-Dependent Dissolution Rate of Nitrofurantoin from Commercial Suspensions, Tablets, and Capsules, Journal of Pharmaceutical Sciences. 63 (1974) 643–645. https://doi.org/10.1002/jps.2600630441.

[26] L. Chen, G. Pilania, R. Batra, T.D. Huan, C. Kim, C. Kuenneth, R. Ramprasad, Polymer informatics: Current status and critical next steps, Materials Science and Engineering: R: Reports. 144 (2021) 100595. https://doi.org/10.1016/j.mser.2020.100595.

[27] S. Langner, F. Häse, J.D. Perea, T. Stubhan, J. Hauch, L.M. Roch, T. Heumueller, A. Aspuru-Guzik, C.J. Brabec, Beyond Ternary OPV: High-Throughput Experimentation and Self-Driving Laboratories Optimize Multicomponent Systems, Adv. Mater. 32 (2020) 1907801. https://doi.org/10.1002/adma.201907801.

[28] B. Burger, P.M. Maffettone, V.V. Gusev, C.M. Aitchison, Y. Bai, X. Wang, X. Li, B.M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R.S. Sprick, A.I. Cooper, A mobile robotic chemist, Nature. 583 (2020) 237–241. https://doi.org/10.1038/s41586-020-2442-2.

[29] A.E. Gongora, B. Xu, W. Perry, C. Okoye, P. Riley, K.G. Reyes, E.F. Morgan, K.A. Brown, A Bayesian experimental autonomous researcher for mechanical design, Sci. Adv. 6 (2020) eaaz1708. https://doi.org/10.1126/sciadv.aaz1708.

[30] B.P. MacLeod, F.G.L. Parlane, T.D. Morrissey, F. Häse, L.M. Roch, K.E. Dettelbach, R. Moreira, L.P.E. Yunker, M.B. Rooney, J.R. Deeth, V. Lai, G.J. Ng, H. Situ, R.H. Zhang, M.S. Elliott, T.H. Haley, D.J. Dvorak, A. Aspuru-Guzik, J.E. Hein, C.P. Berlinguette, Self-driving laboratory for accelerated discovery of thin-film materials, Sci. Adv. 6 (2020) eaaz8867. https://doi.org/10.1126/sciadv.aaz8867.

CRediT Author Statement (Author Contributions):

Pauric Bannigan: Conceptualization, Validation, Investigation, Visualization, Writing-Original draft, Writing-Review and Editing, Project administration. **Florian Häse**: Methodology, Software, Formal analysis, Visualization, Data Curation, Writing-Review and Editing. **Matteo Aldeghi**: Formal analysis, Investigation, Validation, Visualization, Data Curation, Writing-Review and Editing Resources. **Zeqing Bao**: Validation, Visualization, Investigation, Writing-Review and Editing. **Alán Aspuru-Guzik**: Supervision, Writing-Review and Editing, Funding acquisition. **Christine Allen**: Conceptualization, Supervision, Project administration, Writing-Review and Editing, Funding acquisition.