Machine learning modelling in predicting and optimizing PLGA nanoparticle encapsulation efficiency and therapeutic efficacy

by

Sihan Dong

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Science

in

Pharmacy

Waterloo, Ontario, Canada, 2023

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Sihan Dong was the sole author of Chapter 1, 3, 4 which were written under the supervision of Dr. Emmanuel Ho. Exception to sole authorship of material is as follows:

Research presented in Chapter 2:

This research was conducted by Sihan Dong under the supervision of Dr. Emmanuel Ho, Dr. Pascal Poupart, and Dr. William Wong. Sihan Dong designed the experimental studies of this chapter. Wet-lab experiments were carried out by Sihan Dong at the University of Waterloo. The objectives design and a small portion of the programming for Gaussian Processes and Bayesian Optimization models were carried out by Sihan Dong. The vast majority of the programming and model optimization were carried out by Haolin Yu at the University of Waterloo.

Abstract

Nanoparticles (NP) have become a promising drug delivery system in the past few decades in pharmaceutics for its diversity in encapsulating different types of drugs, including proteins/peptides, nucleic acids and small molecule drugs for the treatment of a variety of diseases. Application in cancer cell NP-based drug delivery has been a majority focus because of NP's capability in delivering effective treatment while keeping side effects low. Often, series of chemical and biological assays need to be carried out to pursue certain research goals. However, NP fabrication process is rather time-consuming and costly, consisting of material selection, formation, purification, and characterization.

As NP composition choices can directly influence the NP physicochemical properties and biological behaviors, it is crucial to find the optimized combination efficiently to achieve better NP performances. To ease the burden of conducting experiments manually, collaboration with artificial intelligence (AI) techniques is likely to be a promising choice. Machine learning (ML) as a sub-concept of AI has been a popular tool in many pharmaceutical sciences studies, such as prediction of protein molecular structures, drug discoveries, high throughput screening, and prediction of drug formulation compositions, etc. It has been of researchers' great interests in implementing this emerging technique to a variety of tasks to speed up pharmaceutics development.

In this study, we formulated 32 doxorubicin (DOX) or docetaxel (DTX)-loaded NPs to train and test ML-based Gaussian Processes (GP) models that can estimate the underlying relationships between four NP composition physicochemical properties (e.g., poly (lactic-co-glycolic) acid (PLGA) molecular weight (MW), PLGA lactic acid: glycolic acid (LA/GA) ratio, PLGA: drug weight ratio, and drug lipophilicity) and the corresponding drug EE% and therapeutic efficacy in ovarian cancer cells. No universal relationships between the predictor and response

variables can be concluded. Three GP models including EE% model, DOX NP IC50 model, and DTX NP IC50 model were evaluated for their prediction accuracies that were measured by normalized-RMSE in testing sets. The normalized RMSE are 0.187, 0.296, and 0.206, respectively. The EE% model has the highest prediction accuracy that may be attributed to the larger training dataset compared to the other two models. Furthermore, a simplified Bayesian Optimization (BO) model was built to output a set of x variable values that can potentially help to find formulations that optimize the NP EE% and therapeutic efficacy. In EE% model, the suggested formulation is 2 mg drug with lipophilicity of 2.12 being loaded in 94 mg of 20001 Da, 1.17:1 (LA/GA) PLGA NP. In DOX NP IC50 model, the suggested formulation is 2 mg DOX-loaded 68 mg of 39997 Da, 1.53:1 (LA/GA) PLGA NP. In DTX NP IC50 model, the suggested formulation is 2mg DTX-loaded 90 mg of 20008 Da, 1.70:1 (LA/GA) PLGA NP.

Acknowledgements

First of all, I would like to express my deepest gratitude to my supervisor Dr. Emmanuel Ho for his continuous encouragement, guidance, and generous fundings throughout the years of study. His great mentorship has motivated me to pursue further in research. I would also like to thank my committee members Dr. William Wong and Dr. Pascal Poupart, for giving me valuable advice and support along the journey.

I am especially grateful to my collaborator, Haolin Yu, from Dr. Pascal Poupart's research group. He has been incredibly helpful during our collaboration. Without his expertise in computer science, I would not have been able to complete the project.

I am also grateful to our previous group member Dr. Yannick Traore, for his constant support and help, both academically and personally. I would also like to thank our postdoctoral fellow, Dr. Chen Sun, for his unconditional academic support. Additionally, great appreciation to our School of Pharmacy research technician, Dr. Monica Tudorancea, who has been providing us a safe working environment in the lab. We have also built strong friendship during the years. I am also very fortunate to have our other groups members who feel like families: Jin Wang, Calvin Wong, Alistair Chan, Ethan Watt, Ryan Nodder, and Kaidy Orellena.

At last, I would also like to express my appreciation to Nabeel Tariq, a PhD student from my undergraduate research lab, who has been giving me continuous help and support. Furthermore, I am especially thankful to have a loving family, my parents and grandparents who have never doubted me in accomplishing the program.

Table of Contents

Author's Declaration ii

Statement of Contributions iii

Abstract iv

Acknowledgements vi

List of Figures xi

List of Tables xiv

List of Abbreviations xv

Chapter 1: Introduction 1

1.1 Drug delivery 1

1.2 Polymeric NP formulations 5

1.3 Physicochemical properties of NP formulation material 6

1.3.1 Predictors 7

1.3.1.1 PLGA MW 7

1.3.1.2 PLGA LA/GA ratio 7

1.3.1.3 PLGA: drug weight ratio 8

1.3.1.4 Drug lipophilicity 8

1.3.2 Responses 9

1.3.2.1 EE% 9

1.3.2.2 Half-maximal inhibitory concentration (IC50) 9

1.4 Ovarian cancer (OC) 10

1.5 Chemotherapy drugs 10
1.5.1 DOX 11
1.5.2 DTX 11
1.6 ML applications in pharmaceutics 12
1.6.1 Supervised learning 13
1.6.1.1 GP 13
1.6.1.2 BO 15
1.7 Project rationale, hypothesis, and objectives 16
1.7.1 Project rationale 16
1.7.2 Hypothesis 17
1.7.3 Objectives 17
Chapter 2: Materials and Methods 19
2.1 Materials 19
2.2 PLGA NP formation 19
2.3 PLGA nanoparticle characterization 20
2.4 Indirect drug EE% measurement 20
2.5 Cell culture 21
2.6 MTS cell viability test 21
2.7 GP and BO models 22
Chapter 3: Results and Discussion 24
3.1 PLGA NP 24

3.2 Drug EE% and NP IC50 25
3.2.1 Drug EE% results 28
3.2.1.1 DOX EE% 29
3.2.1.2 DTX EE% 33
3.2.1.3 DOX vs. DTX EE% 37
3.2.2 IC50 38
3.2.2.1 DOX IC50 41
3.2.2.2 DTX IC50 44
3.3 GP model 48
3.3.1 Data preprocessing and split 48
3.3.2 Kernel functions 48
3.3.3 EE% prediction 50
3.3.3.1 Predicted EE% in testing set 50
3.3.3.2 EE% 3D surface response curve 52
3.3.4 IC50 prediction 60
3.3.4.1 Predicted IC50 in testing set 60
3.3.4.2 IC50 3D surface response curve 64
3.4 BO model 68
Chapter 4. Conclusions and future directions 70
4.1 Conclusions 70
4.2 Limitations 73
4.3 Future directions 74

References 76

Appendices 88

Appendix A 88

Appendix B 91

Appendix C 94

Appendix D 97

List of Figures

Figure 1.1. Endocytosis process schematic. Created from BioRender.com. 5

Figure 3.1. DOX UPLC standard curve. Data represent the mean ± SD (N=3). Error bar (standard deviation) is too small to be shown in graph. $R^2 = 0.9999$. 28

Figure 3.2. DTX UPLC standard curve. Data represent the mean ± SD (N=3). Error bar (standard deviation) is too small to be shown in graph. $R^2 = 0.9999$. 28

Figure 3.3. Effect of LA/GA ratio on DOX EE%: A) DOX EE% of 20kDa PLGA NP. B) DOX EE% of 40kDa PLGA NP. Effect of PLGA MW on DOX EE%: C) DOX EE% of 50:50 PLGA NP. D) DOX EE% of 75:25 PLGA NP. **p<0.01, ***p<0.001, ****p<0.0001. Data represent the mean ± SD (N=3). 29

Figure 3.4. Effect of PLGA: drug ratio on DOX EE%: A) DOX EE% of 50:50 20kDa PLGA NP. B) DOX EE% of 50:50 40kDa PLGA NP. C) DOX EE% of 75:25 20kDa PLGA NP. D) DOX EE% of 75:25 40kDa PLGA NP. ***p<0.001, ****p<0.0001. Data represent the mean ± SD (N=3). 32

Figure 3.5. Effect of LA/GA ratio on DTX EE%: A) DTX EE% of 20kDa PLGA NP. B) DTX EE% of 40kDa PLGA NP. Effect of PLGA MW on DTX EE%: C) DTX EE% of 50:50 PLGA NP. D) DTX EE% of 75:25 PLGA NP. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001. Data represent the mean ± SD (N=3). 33

Figure 3.6. Effect of PLGA: drug ratio on DTX EE%: A) DTX EE% of 50:50 20kDa PLGA NP. B) DTX EE% of 50:50 40kDa PLGA NP. C) DTX EE% of 75:25 20kDa PLGA NP. D) DTX EE% of 75:25 40kDa PLGA NP. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001. Data represent the mean ± SD (N=3). 36

Figure 3.7. DTX and DOX EE% at different PLGA MW and LA/GA ratio. A) Drug EE% at 50:50 20kDa PLGA. B) Drug EE% at 50:50 40kDa PLGA. C) Drug EE% at 75:25 20kDa PLGA. D)

Drug EE% at 75:25 40kDa PLGA. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001. Data represent the mean ± SD (N=3). 38

Figure 3.8. Schematic of absolute IC50. 39

Figure 3.9. A) Dose response curve of free DOX drug. IC50 is 0.85□0.23 □g/mL. B) Dose response curve of free DTX drug. IC50 is 0.007□0.002 □g/mL. The free drug cytotoxicity tests follow the same procedure as drug-loaded NP cytotoxicity tests. Data represent the mean ± SD (N=6). Standard deviation error bars that are smaller than the data point symbol are hidden. 40

Figure 3.10. Effect of LA/GA ratio on DOX IC50: A) DOX IC50 of 20kDa PLGA NP. B) DOX IC50 of 40kDa PLGA NP. Effect of PLGA MW on DOX IC50: C) DOX IC50 of 50:50 PLGA NP. D) DOX IC50 of 75:25 PLGA NP. Multiple t test was performed. *p<0.05. Data represent the mean ± SD (N=6). 41

Figure 3.11. Effect of PLGA: drug ratio on DOX IC50: A) DOX IC50 of 50:50 20kDa PLGA NP. B) DOX IC50 of 50:50 40kDa PLGA NP. C) DOX IC50 of 75:25 20kDa PLGA NP. D) DOX IC50 of 75:25 40kDa PLGA NP. One way ANOVA test was performed. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001. Data represent the mean ± SD (N=6). 43

Figure 3.12. Effect of LA/GA ratio on DTX IC50: A) DTX IC50 of 20kDa PLGA NP. B) DTX IC50 of 40kDa PLGA NP. Effect of PLGA MW on DTX IC50: C) DTX IC50 of 50:50 PLGA NP. D) DTX IC50 of 75:25 PLGA NP. Multiple t test was performed. *p<0.05, **p<0.01, ***p<0.001. Data represent the mean ± SD (N=6). 45

Figure 3.13. Effect of PLGA: drug ratio on DTX IC50: A) DTX IC50 of 50:50 20kDa PLGA NP. B) DTX IC50 of 50:50 40kDa PLGA NP. C) DTX IC50 of 75:25 20kDa PLGA NP. D) DTX IC50 of 75:25 40kDa PLGA NP. One way ANOVA test was performed. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001. Data represent the mean ± SD (N=6). 47

Figure 3.14. Regression curve of predicted EE% versus observed EE% in testing set. Regression line function: y = 0.63x+11.68 (R^2= 0.65)..... 50

Figure 3.15. 3D surface response curve for NP EE% at two fixed levels of drug hydrophobicity (logP), LA/GA ratio, polymer: drug ratio, and polymer MW. The fixed levels are the lower and upper limit for each input variable..... 58

Figure 3.16. Regression curve of predicted DOX NP IC50 versus observed DOX NP IC50 in testing set. Regression line function: y = 0.557x+2.512 (R^2 = 0.40)..... 60

Figure 3.17. Regression curve of predicted DTX NP IC50 versus observed DTX NP IC50 in testing set. Regression line function: y = 0.798x+0.00085 (R^2 = 0.66)..... 62

Figure 3.18. 3D surface response curve for DOX NP IC50 at two fixed levels of LA/GA ratio (1, 3), polymer: drug ratio (20, 50), and polymer MW (20000, 40000). 65

Figure 3.19. 3D surface response curve for DTX NP IC50 at two fixed levels of LA/GA ratio (1, 3), polymer: drug ratio (20, 50), and polymer MW (20000, 40000). 67

List of Tables

Table 2.1. Hyperparameters in Matérn kernel. 23

Table 3.1. Summary of values in four x variables (predictor) and two y variables (response) collected from in-vitro experiments. Data represent the mean ± SD (N=3 for EE%, and N=6 for IC50). 25

Table 3.2. Observed and predicted EE% in testing set. Bias is calculated as the absolute value of 100%*(Predicted EE% – Observed EE%)/Observed EE%. 50

Table 3.3. RMSE value for both training and testing models in EE% prediction. 51

Table 3.4. n_restarts_optimizer value and optimized hyperparameters in EE% model. 51

Table 3.5. Observed and predicted DOX NP IC50 in testing set. Bias is calculated as the absolute value of 100%*(Predicted IC50 – Observed IC50)/Observed IC50. 60

Table 3.6. RMSE value for both training and testing models in DOX NP IC50 prediction. 61

Table 3.7. n_restarts_optimizer value and optimized hyperparameters in DOX NP IC50 model. 61

Table 3.8. Observed and predicted DTX NP IC50 in testing set. Bias is calculated as the absolute value of 100%*(Predicted IC50 – Observed IC50)/Observed IC50. 62

Table 3.9. RMSE value for both training and testing models in DTX NP IC50 prediction. 63

Table 3.10. n_restarts_optimizer value and optimized hyperparameters in DTX NP IC50 model. 63

Table 3.11. Summary of BO-suggested x variable values to potentially optimize EE% and IC50. 69

List of Abbreviations

AI	Artificial intelligence
ACN	Acetonitrile
BO	Bayesian optimization
CS	Computer science
DCM	Dichloromethane
DLS	Dynamic light scattering
DOX	Doxorubicin
DOX·HCl	Doxorubicin hydrochloride
DTX	Docetaxel
EDTA	Ethylenediamine tetraacetic acid
EE%	Encapsulation efficiency
ELS	Electrophoretic light scattering
EOC	Epithelial ovarian cancer
EPR	Enhanced permeability and retention
FBS	Fetal bovine serum
FDA	The United States Food and Drug Administration
GI	Gastrointestinal
HGS	High-grade serum
IC50	Half-maximal inhibitory concentration
L-BFGS-B	Limited-memory Broyden-Fletcher-Goldfarb-Shanno Bound
LA/GA	Lactic acid to glycolic acid
LGS	Low-grade serum

logP	The logarithm of partition coefficient
LOO CV	Leave-one-out cross validation
ML	Machine learning
MTS	3-(4,5-dimethylthiazol-2-yl)-5-(3-carboxymethoxyphenyl)-2-(4-sulfophenyl)-2H-tetrazolium, inner salt
MW	Molecular weight
NP	Nanoparticle
OC	Ovarian cancer
P/S	Penicillin/streptomycin
PDA	Photo diode array
PEG	Poly-ethylene glycol
PEI	Poly-ethylene imine
PLGA	Poly (lactic-co-glycolic) acid
PVA	Poly-vinyl alcohol
PVP	Poly-vinyl pyrrolidone
QSAR	Quantitative structure-activity relationships
R^2	Determination coefficient
RBF	Radial basis function
RMSE	Root mean square error
RPMI	Roswell Park Memorial Institute
SD	Standard deviation
UCB	Upper confidence bound
UPLC	Ultra-performance liquid chromatography

W_1/O	Water-in-oil
$W_1/O/W_2$	Water-in-oil-in-water

Chapter 1: Introduction

1.1 Drug delivery

Conventional drug delivery faces multiple barriers that significantly constrains drug distribution and accumulation at target sites and eventually results in reduction of the therapeutic efficacy. Common drug administration routes include intravenous, subcutaneous, intramuscular, oral, topical, and pulmonary, etc[1]. Different from other administrated drugs, oral drugs encounter first pass effects mainly in the liver where drugs are metabolized. Thus, it ends up with a significant reduction in drug bioavailability which accounts for the percentage of drug available in the bloodstream upon initial dosing[2]. In order to be absorbed by the body, oral drugs need to be well-solubilized in different dosage forms and be able to permeate biological barriers such as gastrointestinal (GI) membrane to enter systemic circulation. Hydrophobic drugs that are insoluble in aqueous solutions usually need to be formulated by mixing with extra hydrophilic excipients that may be toxic to humans. On the other hand, these drugs usually have better membrane permeability because of their high lipophilicity and are favored in neurological diseases owing to their abilities of crossing the blood-brain-barrier. However, they experience low bioavailability in systemic circulation after its dissolution in the GI tract. In contrast, hydrophilic drugs experience challenges in crossing the cell membrane but are readily solubilized in the bloodstream and cytoplasm. Besides small molecule drug delivery, gene therapy started its rapid development since it was first delivered to a patient by W. French Anderson in 1990[3]. Unlike small molecule drugs, these genetic therapeutics include DNA and RNA macromolecules that need to be encapsulated in vehicles such as NP and viral delivery vectors in order to enter cells[3,4]. This strategy is also applicable for small molecule drugs to avoid restricted delivery as mentioned above. In these

aspects, NP has become a promising delivery system that improves therapeutics bioavailability and therapeutic effects.

Several types of NP delivery systems have been created to maintain a relatively sustained and controlled drug release in the form of liposomes, micelles, polymeric NP, and metallic NP[5]. The release can be achieved through drug diffusion or NP matrix erosion/degradation depending on the drug physicochemical properties, NP composition, and biological environment the NPs are in[6]. The goals of NPs are to transport sufficient amount of drugs to target sites with advantage of crossing biological barriers, especially the blood-brain-barrier that blocks the entry of drug molecules by tight-junctions[7,8]. And to enhance drug stability by protecting it from enzymatic degradation or hepatic metabolism. NPs are also designed to enhance biocompatibility and to reduce unwanted cytotoxicity. Conventional drugs are free drugs in common forms of oral tablets or injectable solutions with excipients. However, their therapeutic effects are limited due to shortened half-life and insufficient drug accumulation. Regardless of the drug lipophilicity, NP is capable for delivering an extended range of drugs compared to conventional drug deliveries which end up in low bioavailability. NP formulations especially improve the delivery of hydrophobic drug as its low water solubility makes it difficult to be formulated in different dosage forms. Owing to its outstanding merits, NP has been recognized by scientists for its potential to improve drug delivery[9]. NP formulations can be modified to accommodate different biological environments and needs to promote their functions. For example, stabilizers such as poly-ethylene glycol (PEG), poly-vinyl alcohol (PVA), chitosan, poly-ethylene imine (PEI), and poly-vinyl pyrrolidone (PVP) help to stabilize the uniform colloidal suspension in preventing NP agglomeration and degradation[10,11]. Antibodies can also be conjugated to NPs for active cell targeting. They

specifically bind to the targeting receptors on cell surfaces and bring NPs into the cell precisely with enhanced specificity and therapeutic efficacy[9,12].

As one of the most life-threatening diseases, cancer has received great attentions and varieties of therapies have been applied for cure[13,14]. However, anticancer treatments still face challenges of inadequate drug accumulation, side effects on normal cells, and multidrug resistance[13,15]. Owing to the multifunctional characteristics, NPs have been proved to reduce toxicity on healthy cells and enhance prolonged drug release at the tumor site to boost therapeutic efficiency[16,17]. Although many studies have been done to optimize the NP formulations, the underlying relationship between NP compositions and NP behaviors are still ambiguous and the research is time-consuming and costly. Therefore, in this study, we focused on examining the therapeutic effects of different NP formulations on cancer cells and the NP encapsulation efficiency (EE%) which will be discussed later. In NP development, besides NP compositions and structures, the actual environment that drugs will be interacting with should also be examined. It is worth mentioning that tumor microenvironment is very distinct from that in other types of disease. Tumor consists of complex vasculature and stroma that potentially influence the entry and distribution of NPs[18]. The enhanced permeability and retention (EPR) effect in tumor cells allow NPs to undergo easier permeation and tissue retention that facilitate NP reaching its target site[19]. As NPs reach the cells, they usually enter cells through endocytosis under different trafficking pathways depending on their sizes[20–22]. Those that are smaller than 200nm undergo clathrin-dependent, clathrin-independent/dynamin-dependent, clathrin-independent/dynamin-independent, or caveolae endocytosis, while those are larger than 200nm undergo macropinocytosis or phagocytosis if the size exceeds 500nm[21]. In all these pathways, the internalized NPs first reach the early endosome and eventually into the digestive enzyme lysosome[22,23]. Late lysosome digests

and degrades incoming particles by primarily increasing the endosomal acidification down to pH 5.0[24]. To avoid the fate of degradation, it is very important for NPs to escape from the endosome timely before the endosome-lysosome fusion. The four main routes to realize endosomal escape are membrane fusion, osmotic pressure, NP swelling, and membrane destabilization[24]. Several NP surface modifications have been introduced to enhance endosomal escape: peptides, viruses, toxins, chemicals, and polymers[25]. Another crucial factor that controls NP destination in endocytosis is the NP surface charge. NP surface charge is indirectly measured by zeta potential under an applied electric field. Zeta potential is the potential at the slipping plane which locates at the farthest layer from NP and connects to the dispersing aqueous solution[26]. The mechanisms of endosomal escape are different between cationic and anionic NPs. Cationic NPs escape due to proton sponge effect in which protons enter endosome and protonate basic chemical groups on the NP surface, following influx of chloride ions and increased osmotic pressure upon the entry of water molecules, resulting in endosome disruption[24,27–29]. Conversely, anionic NPs can be protonated into cationic NPs in the acidic endosome and break up endosome for drug release[29]. However, NPs can be cytotoxic depending on their surface charges and other physicochemical properties. For the goal of generating a stable and protective vehicle, NP materials should be well chosen upon the charge of the cargo considering the drug-NP interactions[30].

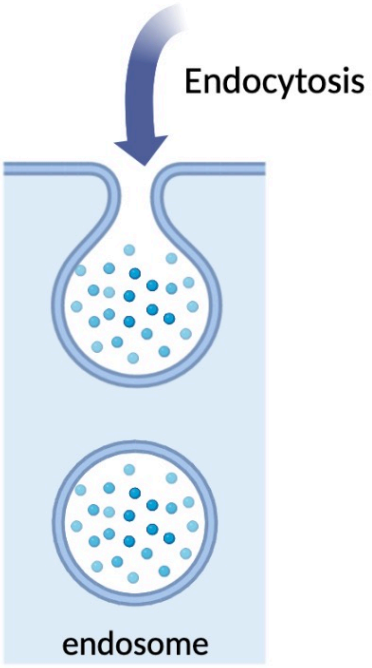


Figure 1.1. Endocytosis process schematic. Created from BioRender.com.

1.2 Polymeric NP formulations

From 2016 to 2021, 29% of United States Food and Drug Administration (FDA) newly-approved NPs are polymeric which is the most common type[31]. However, polymeric NPs only occupied 7% among all the NP clinical trials[31]. There is definitely great potential in polymeric NP development. Polymeric NPs are mainly categorized by their structures: nanospheres, nanocapsules[9,32,33]. The former type is structured with a polymeric matrix, while the later type contains an empty core capsuled by a polymeric shell. Although polymeric NP delivery systems have advantages in controlled release, easy production, and diverse surface modification[34,35], their instability and aggregation remain as concerns in applying clinically[9]. As an external vehicle, NP's fate inside the body is worth to be investigated. NPs face multiple obstacles along the delivery pathway, including immune response, body clearance, blood circulation, biodistribution, and cell

uptake[9]. Cell uptake profile also needed for detailed examination on NP stability and behaviors in a real biological environment. Not only the pharmacokinetics and pharmacodynamics of encapsulated drug need to be studied, that of the NP composition is necessary as well, as it controls the drug-NP interaction and drug release.

To help accelerate polymeric NP development and better predict their EE% and therapeutic efficacies, a set of NPs were formulated in this project for closer examination. We chose to create our own dataset for higher data quality since NP formulations can vary in their compositions and methodologies that are very hard to be kept consistent and comparable among data extracted from literatures.

Poly lactic-co-glycolic acid (PLGA) was chosen to be the main NP component in this project considering its excellent biodegradability, low toxicity, and high biocompatibility[36]. Double emulsion solvent evaporation method was used to fabricate NPs in the form of nanocapsules or nanospheres. Hydrophilic drug was encapsulated inside the water phase while hydrophobic drug was mixed with PLGA in the oil phase. Surfactant PVA serves as stabilizer or emulsifier to form a layer on the NP surface to enhance NP stability. Drug is released either when PLGA degrades through hydrolysis in aqueous medium or through passive diffusion from the NP surface.

1.3 Physicochemical properties of NP formulation material

Material selection of NP and the loading drug greatly impact the NP characterization, physicochemical properties, and cellular behaviors. In polymeric NPs, the choices of polymers determine the NP size, zeta potential, and potentially affect drug release, cellular uptake, and

toxicity[37]. In this study, we mainly investigated the effect of PLGA MW, PLGA LA/GA ratio, PLGA: drug weight ratio, and drug lipophilicity on the NP EE% and therapeutic efficacy.

1.3.1 Predictors

1.3.1.1 PLGA MW

Different polymer MW indicates different polymer chain length and results in different polymer mechanical strength and degradation[38]. Higher MW means longer chains with more ester bonds needed to be broken in hydrolytic degradation and takes up more energy. It also increases polymer hydrophobicity and viscosity that affects drug-polymer interactions. With more interactions, especially hydrophobic drugs tend to retain with the polymer in oil phase. Polymer hydrolysis/degradation speed directly affects the drug release process and how the drug enters cells. Many relevant papers found that NP EE% increases as polymer MW increases because of the reasons above[39–41].

1.3.1.2 PLGA LA/GA ratio

PLGA is composed of two types of repeating units, lactic acid (LA) and glycolic acid (GA). LA with an extra methyl group makes it more hydrophobic than GA. In general, the higher the LA/GA ratio, the more hydrophobic the PLGA. Increased hydrophobicity influences how the polymer interacts with drug and aqueous solution NPs disperse in. This further determines how well drug is encapsulated in NP and how it is released from the NP matrix. Many studies have demonstrated that the NP EE% increased as LA/GA ratio increased due to the increased hydrophobicity and viscosity of PLGA. Drug release is also controlled by the LA/GA ratio-driven PLGA degradation rate and interaction between drugs and both LA and GA monomers. LA/GA

ratio is also responsible for many other PLGA physicochemical properties, including mechanical strength, crystallinity, melting point, and glass transition point[38,42].

1.3.1.3 PLGA: drug weight ratio

The amount of polymers used in NP formulations directly reflects on the NP stability and drug encapsulation capacity by influencing the interactions between PLGA and encapsulated drug, as well as the NP erosion and drug release. The increment of PLGA: drug weight ratio results in an increase of polymer solution viscosity but a decrease in shear stress that is positively proportional to polymer degradation which further controls the drug release rate[40,43].

1.3.1.4 Drug lipophilicity

The logarithm of partition coefficient (logP) which measures the lipophilicity of molecules influences the molecule solubility and the interactions between molecules, NP core matrix (e.g., polymers), and external solution[44–46]. Shown in Equation 1, the higher the logP, the more lipophilic/hydrophobic the molecule. The more similar lipophilicity between the solvent and the drug, the more compatible they are, and higher EE%[47]. This can be explained by the concept of like-dissolves-like which states that molecules with similar polarities tend to dissolve together[48]. This property also impacts the drug release profile: hydrophilic drugs tend to diffuse into the external aqueous media because of its higher solubility in water. In contrast, hydrophobic drugs are prone to interact with the hydrophobic polymer, thus, stay inside the NP matrix. The ways that drugs are released rely on the logP eventually, either through diffusion or happens when NP matrix degrades.

$$\text{logP} = \log\left(\frac{[Solute_{octanol}]}{[Solute_{water}]}\right) \quad (1)$$

1.3.2 Responses

1.3.2.1 EE%

EE% is one of the most important factors in NP development. Higher EE% is demanded but this really depends on the initial amount of drug being encapsulated since the space for drug saturates at a certain drug concentration level. The ability of NP encapsulating small molecule drugs can be varied under different NP formulations, drug properties, and preparation protocols. This includes the polymer, drug, surfactant that were chosen, and the formation methods used. Starting with the same amount of drug, NPs with higher EE is favored to achieve desired therapeutic effects at lower NP dose. In clinical trials, the use of NP with high EE% enhances drug accumulation in cells, helps to reduce side effects on other healthy cells, and may shorten the dosing frequency[16,49,50].

1.3.2.2 Half-maximal inhibitory concentration (IC50)

IC50 is the drug concentration when the drug inhibits 50% of the cell growth. It directly measures the drug potency but also reflects the therapeutic efficacy using a Hill equation[51–53]. The lower the IC50 value is, the more potent and effective the drug is. IC50 also provides an insight into drug efficacy when the therapeutic effects differ between formulations. IC50 relies on many factor in NP: the main one is the drug mechanism of action which is determined by the drug chemical structure and its interactions with the target site. Other factors include the NP formulation components, drug release profile, and cell uptake profile.

1.4 Ovarian cancer (OC)

OC has been one of the major cancer diseases in women and among all types of OC, 90% are epithelial-related[54]. As epithelial ovarian cancer (EOC) is often asymptomatic in early stages, over half of the patients were already at late stages when they were diagnosed[55]. With the urgent needs of effective treatments, different combinatorial therapies have been applied in clinical. Treatments usually consist of debulking surgery and chemotherapy upon the assessment of the stage patients are at[55]. Platinum chemotherapy is a common choice in OC, however, there is a high chance of drug resistance and recurrence after the treatment[56]. Other types of small molecule drugs in such case deserve attention, such as DOX and DTX. Of the two main prototypes of EOC, Type II high-grade serous (HGS) OC, is more aggressive than Type I low-grade serous (LGS) OC and has been the major subtype among diagnosed cases[57]. Herein, human epithelial ovarian cancer cell OVCAR3 from HGS subtype is used as model cell line for further investigation of chemotherapeutic drug-loaded NP to achieve better therapeutic effects.

1.5 Chemotherapy drugs

Most chemotherapy drugs have severe side effects on normal tissue while killing cancer cells and this has arisen as one of the main concerns in anticancer treatments. With the use of NPs, this concern is expected to be relieved by delivering higher amount of small molecule drugs directly to cancer cells by passive targeting facilitated by EPR effect or active targeting through ligand-receptor interactions with minimized cytotoxicity on normal cells[58]. To conduct deeper examinations of NP behaviors, two model drugs are studied for their logP and cytotoxicity in OVCAR3 cells.

1.5.1 DOX

DOX is a very hydrophobic drug due to its chemical structure. Therefore, it is often added with hydrochloride (DOX·HCl) to improve its water solubility in facilitating the delivery into human bodies. To keep it simple, DOX will be used to represent DOX·HCl in the following paragraphs.

DOX has been approved by FDA as intravenous injections in forms of either conventional solutions or liposomal NPs[59]. Adriamycin is the brand name of conventional DOX injection, without drug carriers. However, it can cause severe cardiomyopathy[59,60]. Doxil as the first FDA-approved liposome injection, has been proved to have lower cytotoxicity, increased drug stability, and enhanced therapeutic effects[61]. In terms of mechanism of action, DOX targets DNA, inhibits topoisomerase II which modulates double-stranded DNA entanglement, and generates unstable radicals that cause cell membrane disruption and eventually cell death[62]. DOX has also been a common option in combination with other drugs in OC treatment but has not been as effective[63,64]. Further research should be done on improving the delivery of DOX.

1.5.2 DTX

DTX promotes the tubulin polymerization by interrupting the microtubule polymerization/depolymerization equilibrium and leads to cell death[65,66]. Well-known FDA-approved docetaxel medications include Taxotere and its generic form Docefrez[67]. Both are intravenous injection for administration. DTX has been an effective drug in treating EOC along with platinum drug[63,68] and has also been used in many other cancer diseases, such as breast cancer, lung cancer, and prostate cancer[69]. Studies have been done to show the advantages of DTX-loaded NP delivery, including enhanced drug accumulation in cancer cells and reduced cytotoxicity

systemically[67,70]. However, no DTX-loaded NP formulation has been approved by FDA yet. Further development is necessary to improve the DTX therapeutic efficacy in the form of NPs.

1.6 ML applications in pharmaceutics

ML, one of the subtypes of AI, has expedited and expanded the applications in many fields of study. The applications in healthcare have particularly increased after the corona virus (COVID-19) pandemic in the past few years. ML has also shown the potential in promoting pharmaceuticals development. The power of ML in data utilization has demonstrated its ability in discovering underlying data structures that contain rich information[71]. ML has eased researchers from heavy wet-lab workload, for example, by making prediction, simulating data, and designing experiment[72–74]. The applications have spread among fields including molecular docking, computational biology, drug discovery, drug formulation development, and clinical therapy design[75,76]. In molecular docking, the scoring function that helps assess ligand interaction can be well-generated with higher prediction accuracy upon data training with the use of ML, resulting in the selection of better drug candidates[77,78]. Similarly, with the incorporation of ML, predictions on drug and protein structures, protein functions, genome engineering, and gene expression can also be efficiently improved[79–82]. ML has also been applied to optimize drug dosing and for the prediction of patient responses to treatment in order to design better clinical treatment[76,83].

ML offers numerous choices of algorithms for different data types and analysis purposes. It is mainly categorized into supervised learning, unsupervised learning, and reinforcement learning, according to the data type and ways of model-learning[72]. Supervised learning requires labelled output data that are clearly specified, while output data in unsupervised learning are not labelled. In other words, the target data or patterns in unsupervised learning are undetermined and

it is the job for ML model to discover. Different from supervised and unsupervised learning, reinforcement learning does not require data training, but rather makes decisions to obtain higher rewards upon an environment-interactive action taken by an agent at a certain state. ML has been mostly involved in common tasks of regression and classification in supervised learning, clustering and data dimensionality reduction in unsupervised learning[84]. Popular algorithms include linear regression, artificial neural network, support vector machine, random forests, k-nearest neighbors, principal component analysis, k-means clustering, and so on[85,86]. Each of them models data differently, even though some may from the same type of learning style. Thus, it is very essential to pick the useful algorithm to ensure reliable data fitting and fulfill goals in each project.

In cooperation with ML, NP formulation design can be efficiently improved. For the purposes of making predictions of output values (EE% and IC50), supervised learning models are used in this project.

1.6.1 Supervised learning

1.6.1.1 GP

GP is a non-parametric stochastic probabilistic model which finds the joint distribution over functions and is defined by a mean function $m(\boldsymbol{x})$ and covariance (i.e., kernel) function $k(\boldsymbol{x}, \boldsymbol{x})$ in Equation 2, where the mean function is the expectation of the true function and the covariance function finds the similarity between two inputs. The true function that GP is estimating maps input variables to output variables. As a probabilistic model, GP not only makes predictions on the output values, but also provides estimation of the function distribution uncertainty with the use of covariance function. In real cases, GP can also incorporate input and output data uncertainties[87,88]. As shown in Equation 3, the true function $f(\boldsymbol{x})$ is added with the noise ϵ. GP

follows multivariate (joint) normal distribution which is saying that any set of finite number of random variables follows multivariate normal distribution. GP is based on Bayesian distribution to update a posterior distribution from a prior distribution after data training which is called conditioning. The prior distribution is of the early known information of the model itself, while the posterior distribution is used to make predictions on new data points and estimate prediction uncertainties, given a set of training inputs and outputs. In the posterior predictive distribution (Equation 4), $\boldsymbol{x}^*$ represents the testing input, y^* represents the testing output, $\boldsymbol{X}$ and $\boldsymbol{y}$ represent the training input and output, σ^2 represents the error variance.

The hypothesis space of GP consists of all possible functions that can fit the data. These functions are determined by the prior distribution[89]. Kernel matrix has $n \times n$ dimension, where n is the training dataset size. The basis function embedded in kernel functions are able to effectively project x variables into a new feature space and the computational complexity $O(n^3)$ now only depends on n [90]. In the new feature space, data structure is comprehensively examined and the underlying relationships between data points are better revealed. This method speeds up the analyzing process and enhances the prediction accuracy in small dataset scenarios.

$$f(\boldsymbol{x}) \sim GP(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x})) \tag{2}$$

$$y = f(\boldsymbol{x}) + \epsilon \tag{3}$$

$$\mathrm{p}(y^* | \boldsymbol{x}^*, \boldsymbol{X}, \boldsymbol{y}) = N(m'(\boldsymbol{x}^*), \sigma^2 + k'(\boldsymbol{x}^*, \boldsymbol{x}^*)) \tag{4}$$

Kernel functions are mainly categorized into stationary and non-stationary kernels. Stationary kernels are determined by the differences between two sample points, rather than their locations[91,92]. Stationary kernels are subdivided into anisotropic (which considers both distance and direction between two points) and isotropic (only considers the distance between two points) kernels[91]. In contrast, non-stationary kernels focus on the locations of two sample points in the

input space[91,92]. In many cases, non-stationary kernels are more flexible to model data with complex structure but also require more parameter optimizations. There are many types of kernel functions available for GP: radial basis function (RBF) (Equation 5), Matérn kernel (Equation 6), periodic kernel, linear kernel, rational quadratic kernel, and dot product kernel[93]. In real world cases, a combination of kernels may be necessary to obtain a model that best explains the data. The combination can be multiplication and addition of multiple kernels.

The most commonly used kernel is RBF which is also named as Gaussian kernel or squared exponential kernel (Equation 5). $\boldsymbol{x}$ and $\boldsymbol{x}'$ are two different input variables, σ^2 is the variance, and l is the length scale that controls the function smoothness. The smaller the l is, the easier the prior function changes with x, the more fluctuating the function is[93–95].

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 e^{\left(-\frac{||x-x'||^2}{2l^2}\right)} \tag{5}$$

In multivariate regression cases, Matérn kernel (Equation 6) is commonly applied, where v determines the smoothness of the function, d accounts for the distance between two inputs, Γ represents the gamma function, and K_v represents the modified Bessel function[93,95,96].

$$k(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{\Gamma(v)2^{v-1}}\left(\frac{\sqrt{2v}}{l} d(\boldsymbol{x}, \boldsymbol{x}')\right)^{v} K_v\left(\frac{\sqrt{2v}}{l} d(\boldsymbol{x}, \boldsymbol{x}')\right) \tag{6}$$

1.6.1.2 BO

BO is a popular optimization method to find the best combination of input variables that maximize or minimize an objective function which is the function of interest. In this project, it is the function between the four predictors and two responses mentioned before. To reduce the computational cost, a surrogate model with a selected number of data is used to build the relationship between inputs and outputs, approximating the true objective function. One of the

most popular choices of surrogate model is GP. GP is a non-parametric model that is not constrained to a fixed number of parameters or functional form, so that it is more flexible in modeling data. It provides BO an estimated objective function with a mean and covariance function.

The other component in BO is the acquisition function which suggests the next x value that's worth to be examined in achieving higher EE% or lower IC50. It provides suggestions upon an agreement between balancing exploitation and exploration. In exploitation, the acquisition function suggests to evaluate areas around the current best x values with low risk. In contrast, in exploration the acquisition function suggests x values from unexplored region. A new set of x values will be created after each iteration in BO. The corresponding output y values should be collected by conducting wet-lab experiment and used to update GP model to be prepared for the next optimization iteration.

1.7 Project rationale, hypothesis, and objectives

1.7.1 Project rationale

NP as a functional drug delivery platform faces challenges in selecting materials that are beneficial in improving its biological behaviors to meet specific criteria. The fabrication process of NP is time and money-consuming and the choices of materials directly influence the downstream cellular applications regarding the encapsulation capacity and the release of the cargo. Plenty of NP physicochemical properties are responsible for the NP performance, but the effects of a combination of them are still unseen. Two of the main expectations of NPs are to achieve higher EE% and lower IC50 to enhance NP therapeutic efficacy and potency in delivering more drugs to the target site while minimizing the side effects on normal cells. However, the

optimization is not as efficient in *in-vitro* lab work, as experimental design is built on the basis of controlling variables that require a significant amount of work to accurately describe the contribution of each variable to the responses. Beyond this, it is also difficult to find the contribution of a combination of variables without mathematical analysis.

As ML technique emerged rapidly in the past decade, its advantage in accelerating drug design and development drew our attention. With the leverage of ML-based models, the black box of the relationships between NP composition properties and NP performances can be disclosed by training a subset of *in-vitro* data that is composed of input (material physicochemical properties) and output (EE% and IC50) variables at different values. The EE% and IC50 can therefore be predicted according to the selected inputs, following input optimization to enhance EE% and reduce IC50.

1.7.2 Hypothesis

We hypothesize that GP and BO models can make accurate predictions and optimizations on NP compositions in achieving high EE% and therapeutic efficacy based on the physicochemical properties of the selected materials.

1.7.3 Objectives

1. Formulate thirty-two DOX or DTX loaded NPs as well as corresponding blank NPs that are composed of different drug and polymers that have four physicochemical properties. Characterize NPs, measure drug EE%, and obtain IC50 value from MTS cell viability test.
2. Analyze how the collected data are influenced by the four drug and polymer physicochemical properties.

3. Build GP and BO model and train them with the collected EE% and IC50 data from objective 1.
4. Tune GP and BO model hyperparameters to improve prediction accuracy.

Chapter 2: Materials and Methods

2.1 Materials

Avanti J-E BioSafe centrifuge and Optima MAX-XP ultracentrifuge were purchased from Beckman Coulter. Q125 sonicator was purchased from Qsonica. FreezeZone 2.5 liter benchtop freeze dry system was purchased from Labconco. Zetasizer Ultra was purchased from Malvern Panalytical. AQUITY ultraperformance liquid chromatography (UPLC) H-Class, HSS C18 column (100 Å, 1.8 μm, 2.1 mm), HSS-C18 VanGuard column (100 Å, 1.8 μm, 2.1 mm×5 mm) were purchased from Waters. Varioskan LUX multimode microplate reader was purchased from ThermoFisher Scientific.

Acid terminated-PLGA were purchased from Nanosoft Polymers. CellTiter 96® AQueous One Solution Cell Proliferation Assay (MTS) was purchased from Promega. OVCAR3 cell line was purchased from ATCC. RPMI 1640 with 2.5 mM L-Glutamine, penicillin/streptomycin 100x (P/S), heat-inactivated fetal bovine serum (FBS), 0.4% Trypan Blue were purchased from Fisher Scientific. Insulin from bovine pancreas, dichloromethane (DCM), acetonitrile (ACN), formic acid (FA), and PVA (MW 31,000-50,000 Da, 87-89% hydrolyzed) were purchased from Sigma Aldrich. 0.25% trypsin-EDTA was purchased from ThermoFisher Scientific. DTX and DOX were purchased from TCI Chemicals.

2.2 PLGA NP formation

Following the designated nanoparticle formulations, certain amount of PLGA was dissolved in 2 mL DCM. 2 mg of DOX was dissolved in 200 μL Milli-Q water while 2 mg DTX was dissolved together with PLGA in DCM. The drug amount was kept the same at 2 mg for both types of drugs to ensure the NP EE% are comparable between formulations. 200 μL of 10 mg/mL

DOX was emulsified with the organic phase of PLGA on ice using a microprobe sonicator at 60% amplitude for 20s per cycle for 5 cycles, with 5s of rest between each cycle to form the first water-in-oil (W_1/O) single emulsion. In DTX NP formulation, 200 μL water was used as the first water phase instead. The W_1/O emulsion was then dropwise added to 10 mL 3% (w/v) PVA solution and underwent the same sonication procedure on ice to form the water-in-oil-in-water ($W_1/O/W_2$) double emulsion. The final emulsion was then added with 30 mL 3% (w/v) PVA solution and stirred at 600 rpm for 3 hrs to evaporate the organic solvent. Blank NPs were made under the same double emulsion protocol except that no drug was loaded. NPs were washed twice with doubled volume of water to remove free polymers and drugs using the floor centrifuge. The first supernatant was saved for EE% determination. NPs were washed at 20,000 ×g, 20 °C for 15 mins. Part of the NPs was frozen in -20 °C freezer overnight following freeze-drying for 24 hrs in the freeze dryer. The dried NPs were weighed to be used in NP weight concentration calculation.

2.3 PLGA nanoparticle characterization

NPs were washed and dispersed in MilliQ water for characterization using the Zetasizer. Both size and zeta potential were measured using cuvette DTS1070 in triplicates. NP size was measured with dynamic light scattering (DLS) back scatter technique and zeta potential was measured with electrophoretic light scattering (ELS).

2.4 Indirect drug EE% measurement

DOX was dissolved in MilliQ water and a serial dilution was performed for the standard curve. Similarly, a serial dilution of DTX was prepared in ACN for standard curve measurement. All standard drug solutions were filtered with 0.1 μm filter. The drug concentration in NP

supernatant was tested by photo diode array (PDA) detector, using reversed phase UPLC with a HSS C18 column (100 Å, 1.8 μm, 2.1 mm) which was equipped with HSS-C18 VanGuard column (100 Å, 1.8 μm, 2.1 mm × 5 mm) to avoid contamination. The supernatants collected from NP preparation step were ultracentrifuged at 100,000 ×g for 30min at 25 °C to separate free polymers and remaining NPs from free drug, following filtration with 0.1 μm filter. Filtered DTX samples were detected at 230 nm in mobile phase of ACN: H2O (50%: 50%, v/v) with flow rate 0.4 mL/min at 25 °C. Filtered DOX samples were detected at 233 nm in mobile phase of (A) 0.1% formic acid in H2O: ACN (50%: 50%, v/v) and (B) H2O at a ratio of 65%: 35% (v/v) with flow rate 0.3 mL/min at 25 °C. Mobile phases were filtered by 0.22 μm filter and the sample injection volume is 3 μL. Each sample was measured in triplicates.

$$\text{EE\%} = \frac{Total \text{ amount of } drugs - drugs \text{ } in \text{ the } supernatant}{Total \text{ amount of } drugs} \times 100\% \quad (7)$$

2.5 Cell culture

OVCAR3 cells were cultured in RPMI 1640 with L-Glutamine supplemented with 20% FBS, 1% P/S, and 0.01 mg/mL bovine pancreas insulin in 37 °C, 5% CO_2 cell incubator. Cells were subcultured at 80% confluency. 0.25% trypsin-EDTA was used to detach the cells from cell culture flask following centrifugation at 125 ×g for 5 min. Cryoprotectant of 5% DMSO in complete growth medium was used to freeze cells that were then stored in liquid nitrogen tank.

2.6 MTS cell viability test

MTS cell viability assays were performed to find the IC50 value for free DOX, free DTX, and each NP formulation in OVCAR3 cells. Cells were counted by diluting 100 μL stock cells

solution with 100 μL Trypan Blue[97]. 100 μL cells were seeded in sterile 96-well plate at a density of 5000 cells/well and incubated in 37 °C, 5% CO_2 cell incubator for 24 hrs. On the treatment day, all groups except for blank control group were added with 100 μL additional medium containing different components. In treatment groups, NPs at different concentrations in complete growth medium were added to cells. The positive control group was treated with 5% DMSO in complete growth medium and negative control group was treated with complete cell medium only. Each group consists of 6 replicated wells. Cells were washed with PBS twice to remove the residual NP after 48 hrs of NP treatment. 20 μL MTS reagent with 100 μL complete growth media were added to all groups (including a blank control group) away from light and incubated for 2 hrs before read at wavelength 490 nm using the microplate reader. Free drug IC50 values were obtained following the same protocol as NPs with free drugs in treatment groups instead.

MTS validation test was performed by seeding cells at different densities and equilibrated for 1.5 hrs. 20 μL MTS reagent was then added to each well and incubated for 2 hrs in 37°C, 5% CO2 cell incubator. All absorbances were subtracted by the average reading of blank group (n=6) which only contains MTS reagent and medium. Results are shown in Appendix B Figure B1.

$$\text{Cell viability} = \frac{Treatment\ group - Average\ blank\ group}{Average\ negative\ control\ group - Average\ blank\ group} \times 100\% \quad (8)$$

2.7 GP and BO models

GP prior distribution was defined by prior mean of 0 and Matérn kernel function. GaussianProcessRegressor function was specified with a default 'optimizer' of 'fmin_l_bfgs_b' which is named Limited-memory Broyden-Fletcher-Goldfarb-Shanno Bound (L-BFGS-B). The optimizer minimizes the negative log marginal likelihood function to tune the hyperparameters.

Another parameter 'n_restarts_optimizer' was set to 20. The 'length_scale', 'length_scale_bounds', and 'nu' hyperparameters in Matérn kernel were set as follows.

Table 2.1. Hyperparameters in Matérn kernel.

	Prior mean	**length_scale**	**length_scale_bounds**	**nu**
EE%	0	1	0.1, 30	0.5
DOX IC50		1	0.1, 30	0.5
DTX IC50		1	2, 30	0.5

BO utilizes GP as the surrogate function to predict the objective function which was set to 'none'. The acquisition function was set by "UtilityFunction' using upper confidence bound (UCB) in which the hyperparameter 'kappa' controls the trade-off between exploration and exploitation. 'kappa' was set to 0 since no extra experiments were performed for GP model update.

Input data and code for GP and BO models are presented in supplementary files in Borealis repository.

Chapter 3: Results and Discussion

3.1 PLGA NP

Bilati et al. has found a decreasing trend on NP size as sonication amplitude increases before reaching a plateau[98]. This trend also stands for sonication time tests[98]. Sonication intentionally helps to break up aggregated NP, however, high sonication amplitude and long sonication time also have chance to introduce disturbance to NP, resulting in aggregation. Higher sonication amplitude provides stronger power which intensifies NP movements and increases NP collision frequency. The amplitude was optimized to 60% for NP formation in this study.

We found that NP is only stable (in terms of precipitation-free) within a certain range of PLGA concentration but starts to lose the Tyndall effect and fails to be resuspended beyond lower and upper limits of PLGA concentration. Viscosity of PLGA in DCM solution increases as its concentration increases and the concentration should be kept lower than 50mg/mL for better practice. Therefore, the PLGA: drug weight ratio range was selected to fall between 20:1 and 50:1, where the amount of drug was fixed at 2 mg.

Serving as a surfactant, PVA helps to stabilize the NP in avoiding coalescences and the NP stability increases with PVA concentration. The addition of PVA can effectively reduce the Gibbs free energy and interfacial tension to maintain a spontaneous formation of emulsion droplets, especially at higher concentrations[40,99,100]. Research has found that less hydrolyzed PVA helps to form more stable emulsion droplets, and this was confirmed by gelatinization studies where the organic solvent acetone is added to PVA solution to form precipitated PVA gels[101]. At lower degree of hydrolysis, PVA has less hydroxyl groups to form hydrogen bonds between or within PVA molecules, thus, less aggregation and more interactions with water molecules to stabilize NPs[101]. Herein, we used 87-89% hydrolyzed PVA to form well-dispersed NP.

Blank NPs cytotoxicity studies were performed following the same protocol as drug-loaded NPs. Data are shown in Appendix D Figure D1 and D2. Blank NPs composed with 50:50 20kDa/40kDa and 75:25 20kDa/40kDa PLGA at 20:1 and 50:1 PLGA: drug ratio were produced. The cytotoxicity of NPs at 30:1 and 40:1 PLGA: drug ratio were extrapolated from the above-mentioned NPs. That is, for example, if NPs at 20:1 and 50:1 PLGA: drug ratio both are non-toxic at 100 μg/mL, then NPs at 30:1 and 40:1 PLGA: drug ratio are also non-toxic at or below this concentration. Some of the NPs at 30:1 and 40:1 PLGA: drug ratio were also tested in cases that extrapolation failed.

The size and zeta potential data of all formulated blank and drug-loaded NPs are listed in Appendix A Table A1.

3.2 Drug EE% and NP IC50

Table 3.1. Summary of values in four x variables (predictor) and two y variables (response) collected from in-vitro experiments. Data represent the mean ± SD (N=3 for EE%, and N=6 for IC50).

logP	PLGA MW (Da)	PLGA: drug weight ratio	LA/GA ratio	EE%	IC50 (μg/mL)
	20k	20:1	50:50	20.89±0.04	4.04±1.72
		30:1		27.92±0.23	3.78±1.61
		40:1		35.35±0.17	2.44±1.25
		50:1		48.16±0.06	1.60±0.55
		20:1		20.01±0.24	4.62±1.44

	40k	30:1	75:25	21.48±0.32	4.34±2.45
		40:1		33.96±0.03	3.70±1.91
		50:1		38.02±0.12	1.13±0.41
0.36 (DOX)	20k	20:1		30.58±0.17	8.70±3.77
		30:1		35.53±0.03	7.85±3.64
		40:1		37.74±0.07	2.61±1.53
		50:1		45.50±0.15	2.38±0.81
	40k	20:1		16.34±0.31	11.85±5.26
		30:1		17.47±0.09	3.67±1.74
		40:1		21.73±0.19	3.06±0.82
		50:1		28.46±0.17	2.16±0.52
2.74 (DTX)	20k	20:1	50:50	24.59±0.23	0.0187±0.0032
		30:1		38.10±0.25	0.0116±0.0027
		40:1		42.63±0.53	0.0064±0.0015
		50:1		40.70±1.19	0.0019±0.0005
	40k	20:1		34.59±0.21	0.0199±0.0053
		30:1		36.42±0.16	0.0084±0.0024
		40:1		42.46±0.77	0.0089±0.0034
		50:1		45.32±0.40	0.0072±0.0017
		20:1		31.49±0.10	0.0362±0.0139

20k	30:1	75:25		38.55±0.32	0.0069±0.0032
	40:1			45.97±0.63	0.0123±0.0030
	50:1			45.91±0.08	0.0068±0.0020
40k	20:1			24.96±0.28	0.0205±0.0043
	30:1			37.63±0.57	0.0143±0.0038
	40:1			43.64±0.59	0.0129±0.0026
	50:1			47.71±0.08	0.0122±0.0032

Table 1 listed the EE% and IC50 values for all drug-loaded NPs with different drugs and at different PLGA MW, PLGA: drug ratio, and PLGA LA/GA ratio. The drug logP was calculated from the consensus logP from the SwissADME software[102]. For convenience, formulations will be denoted in the form of "PLGA: drug ratio_LA_PLGA MW". For example, 20_50_20k is NP at 20:1 PLGA: drug ratio with 50:50 20kDa PLGA. The relationships between the four drug and PLGA physicochemical properties will be discussed in later sections.

3.2.1 Drug EE% results

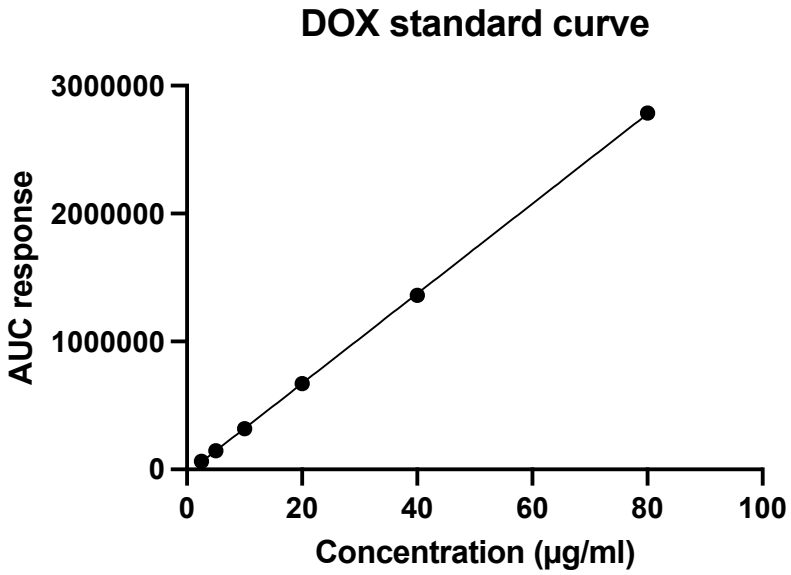


Figure 3.1. DOX UPLC standard curve. Data represent the mean ± SD (N=3). Error bar (standard deviation) is too small to be shown in graph. $R^2 = 0.9999$.

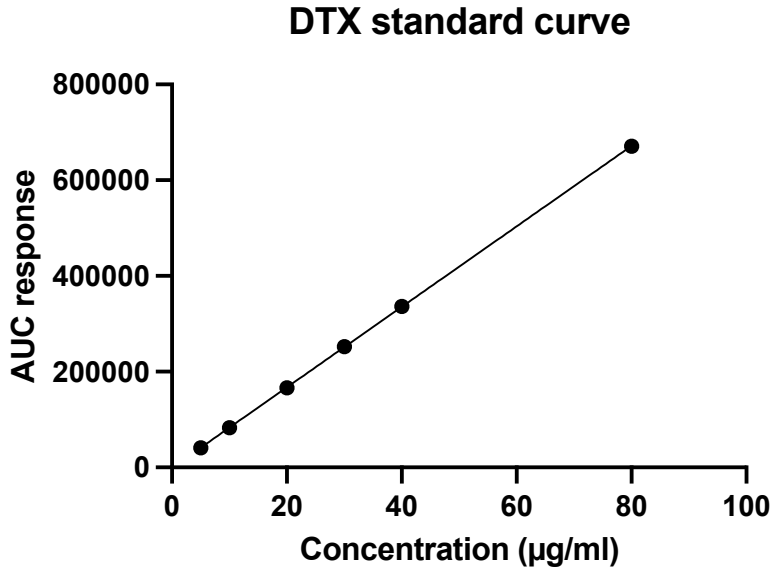
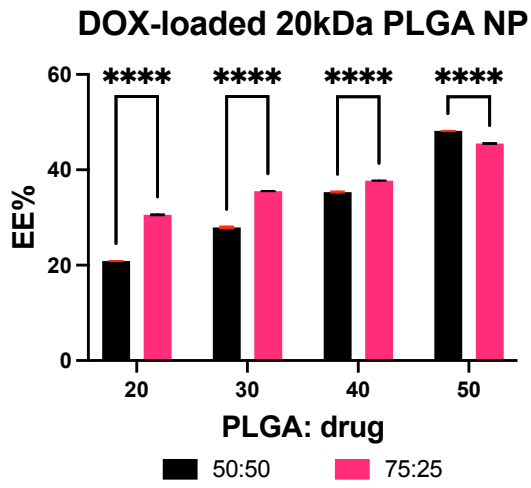


Figure 3.2. DTX UPLC standard curve. Data represent the mean ± SD (N=3). Error bar (standard deviation) is too small to be shown in graph. $R^2 = 0.9999$.

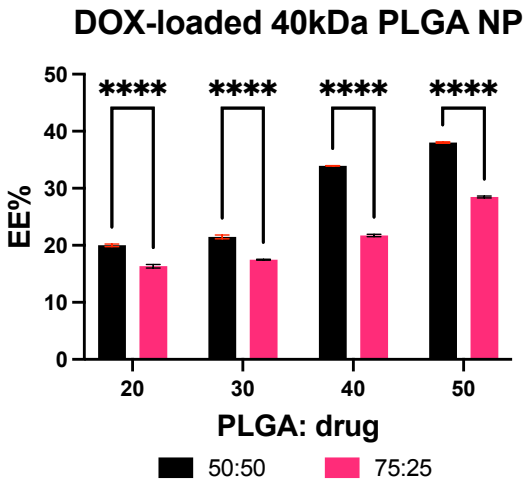
The UPLC standard curves of DOX and DTX follow a linear regression relationship with $R^2 = 0.9999$ (Figure 3.1 and 3.2), indicating a high detection accuracy in UPLC. All drug EE% were calculated based on the standard curves.

3.2.1.1 DOX EE%

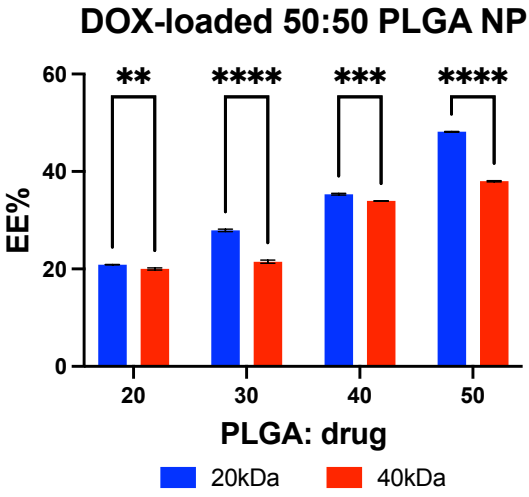
A.



B.



C.



D.

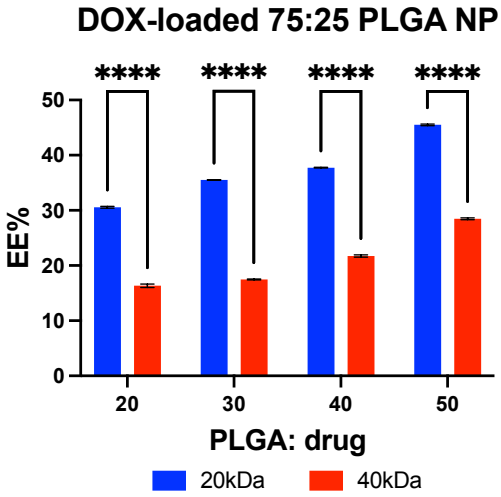


Figure 3.3. Effect of LA/GA ratio on DOX EE%: A) DOX EE% of 20kDa PLGA NP. B) DOX EE% of 40kDa PLGA NP. Effect of PLGA MW on DOX EE%: C) DOX EE% of 50:50 PLGA

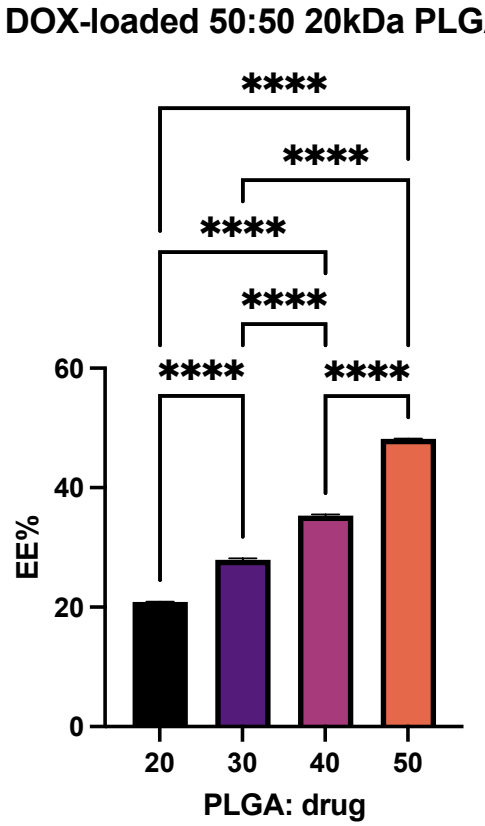
NP. D) DOX EE% of 75:25 PLGA NP. **p<0.01, ***p<0.001, ****p<0.0001. Data represent the mean ± SD (N=3).

Typically, the LA repeating unit in PLGA is more hydrophobic than GA repeating unit. Hence, the PLGA hydrophobicity at 75:25 LA/GA ratio is higher than that at 50:50 LA/GA ratio. According to the molecular dynamic simulation model produced by Stipa's group, LA unit tends to keep its movement within itself (i.e., more self-interaction), while in the presence of both LA and GA units, PLGA interacts with the surrounding aqueous solvent more[103]. In the case of hydrophilic drug encapsulation, the drug-polymer interaction decreases as polymer hydrophobicity increases. The drug diffuses easier into the external aqueous phase, and results in decreased NP EE%. This explains the result in Figure 3.3B, where DOX EE% is lower in 75:25 40kDa PLGA NP, due to the increased PLGA hydrophobicity. However, when PLGA MW is at 20kDa in Figure 3.3A, DOX EE% generally increases with PLGA LA/GA ratio, other than 50_75_20k NP. The inverse relationship might be due to the decreased PLGA self-interaction due to reduced chain entanglement at lower MW[104]. It is likely that DOX has more interaction with PLGA in this case, so that it may diffuse into the continuous external phase with low MW PLGA, especially when LA/GA ratio is also low (e.g., 50:50). Drug-polymer interaction plays the main role when PLGA MW is low, while PLGA self-interaction is dominant when PLGA MW is high.

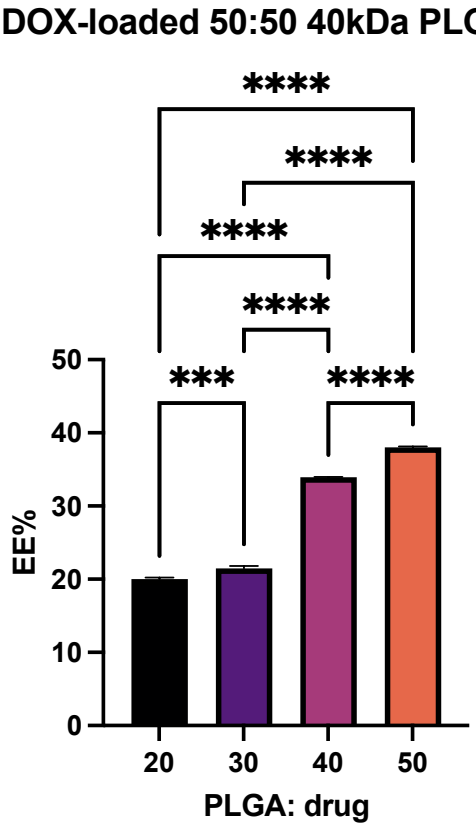
The effects of PLGA self-interaction and drug-PLGA interaction at different PLGA MW on DOX EE% are reflected in Figure 3.3C and D. DOX EE% of 20kDa PLGA NP are higher than that of 40kDa PLGA NP, owing to increased drug-PLGA interactions. Similarly, Presmanes et al. observed the same trend in PLGA NPs loaded with hydrophilic protein[105]. Conversely, some research papers found that hydrophilic drug EE% increases as PLGA MW increases, regardless of

the NP fabrication methods (single emulsion and double emulsion)[40,41,106]. There is lack of evidence on how exactly drug EE% changes with PLGA MW.

A.



B.



C.

D.

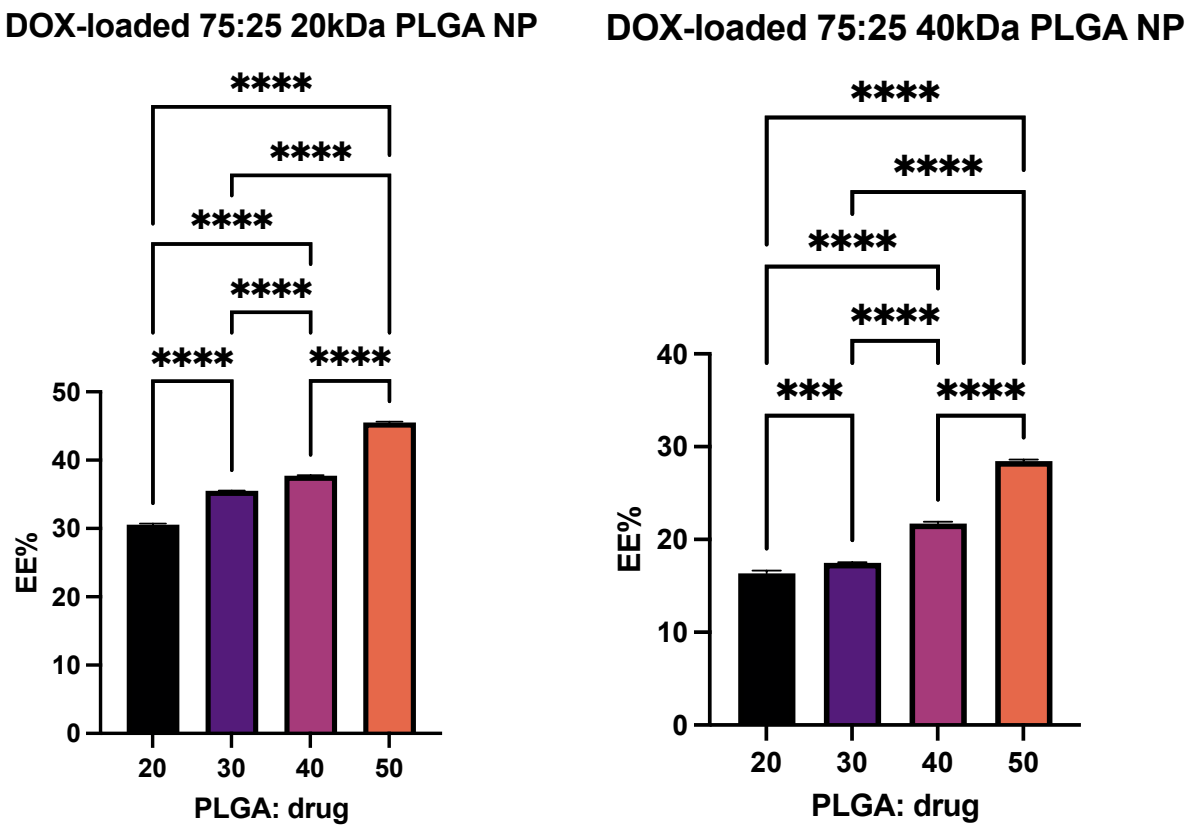
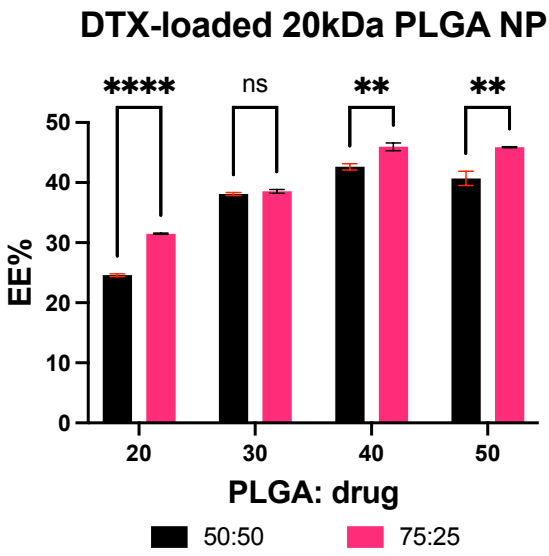


Figure 3.4. Effect of PLGA: drug ratio on DOX EE%: A) DOX EE% of 50:50 20kDa PLGA NP. B) DOX EE% of 50:50 40kDa PLGA NP. C) DOX EE% of 75:25 20kDa PLGA NP. D) DOX EE% of 75:25 40kDa PLGA NP. ***$p<0.001$, ****$p<0.0001$. Data represent the mean ± SD (N=3).

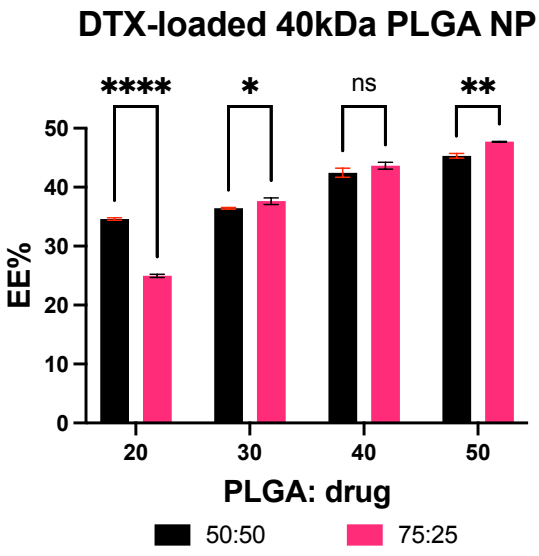
In Figure 3.4, EE% increases as PLGA: drug ratio increases. As PLGA concentration (i.e., PLGA: drug ratio) increases, there are more PLGA chains present in the organic solution for NP construction, so that the solution viscosity increases. This restricts the drug from diffusing into the double emulsion NP continuous aqueous phase. Thus, more drugs were entrapped inside the NP and resulted in increased EE%.

3.2.1.2 DTX EE%

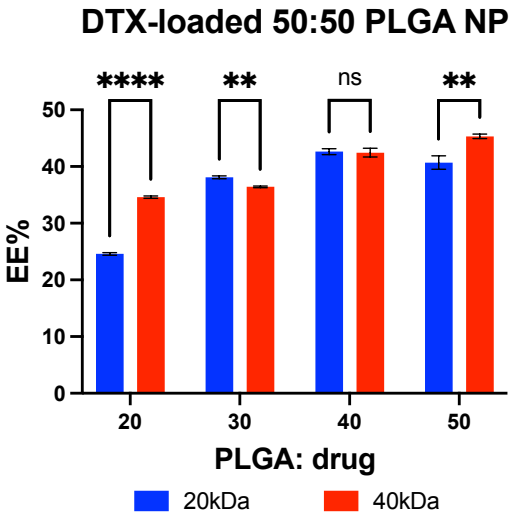
A.



B.



C.



D.

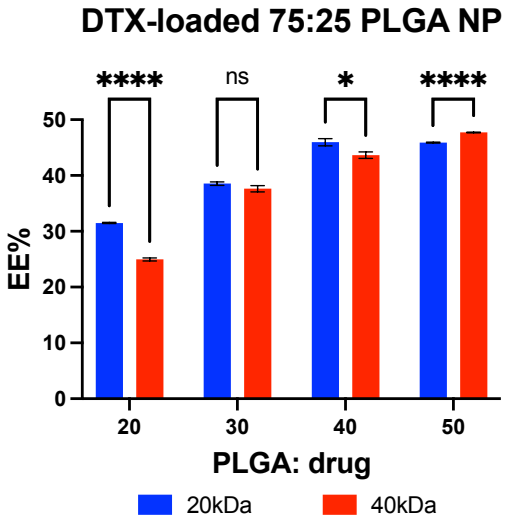
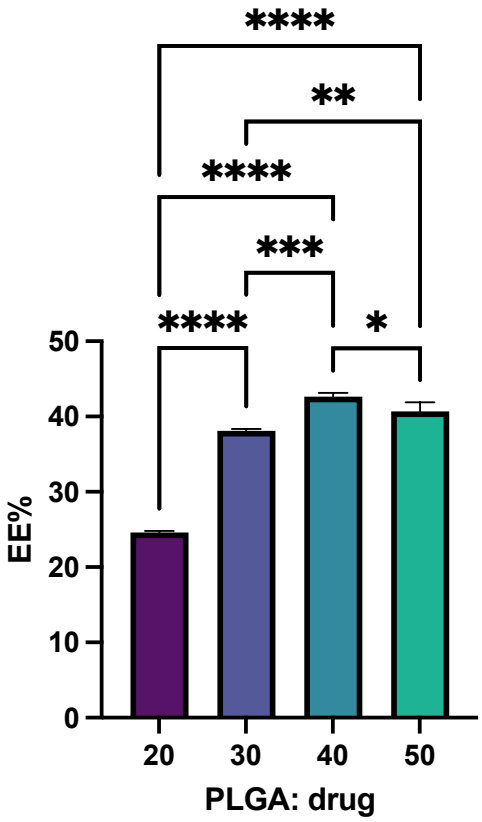


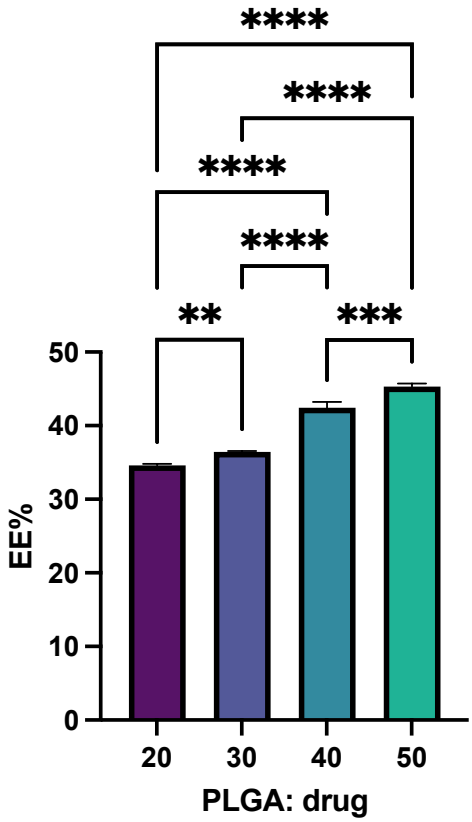
Figure 3.5. Effect of LA/GA ratio on DTX EE%: A) DTX EE% of 20kDa PLGA NP. B) DTX EE% of 40kDa PLGA NP. Effect of PLGA MW on DTX EE%: C) DTX EE% of 50:50 PLGA NP. D) DTX EE% of 75:25 PLGA NP. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001. Data represent the mean ± SD (N=3).

DTX EE% in 75:25 PLGA NP are mostly higher than that of 50:50 PLGA NP when PLGA MW is at 20kDa (Figure 3.5A), except for the 30:1 PLGA: drug ratio NPs that have insignificant EE% difference between each other. The increasing trend can be explained by the fact that DTX interacts more with hydrophobic 75:25 PLGA compared to more hydrophilic 50:50 PLGA. The increased drug-PLGA interaction helps to preserve DTX in oil phase. No clear relationship can be concluded between DTX EE% and PLGA LA/GA ratio in Figure 3.5B when PLGA MW is at 40kDa. The impact of different LA/GA ratio on NP EE% at high PLGA MW 40kDa is too small to be observed. Unlike DOX NP, there is no clear trend between DTX EE% and PLGA MW in Figure 3.5C and D. The difference between the two selected PLGA MW is not significant enough to reflect on DTX EE%. Similarly, Fonseca et al. did not find a clear relationship between hydrophobic drug EE% and PLGA MW, either[107].

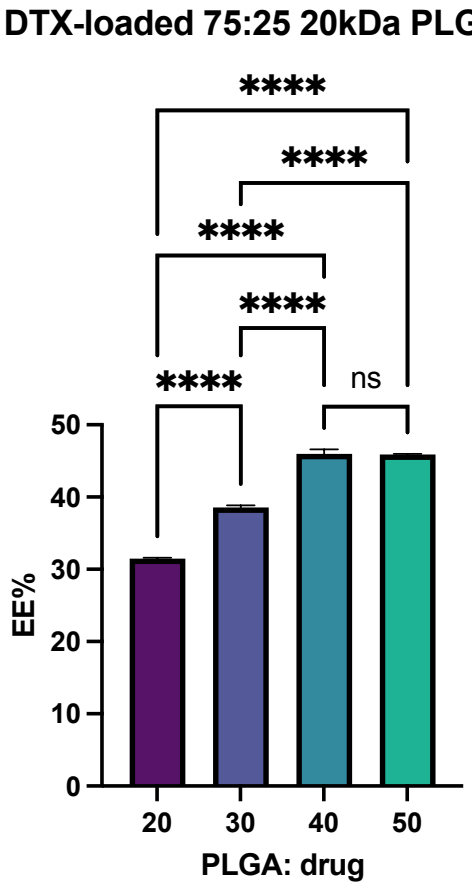
A.


DTX-loaded 50:50 20kDa PLGA NP
****
**
****
***
****
*
50
40
30
20
10
0
EE%
20
30
40
50
PLGA: drug

B.


DTX-loaded 50:50 40kDa PLGA NP
****
****
****
****
**
***
50
40
30
20
10
0
EE%
20
30
40
50
PLGA: drug

C.



D.

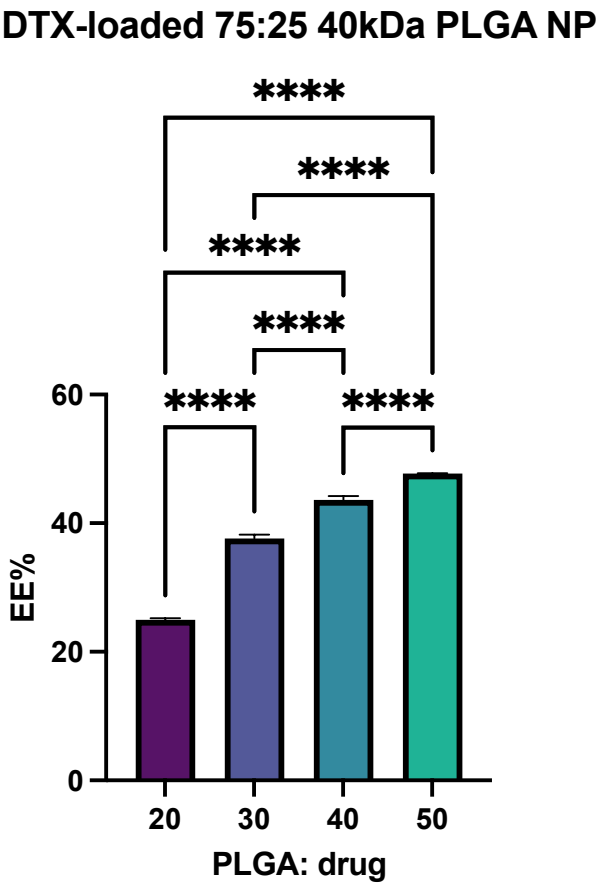


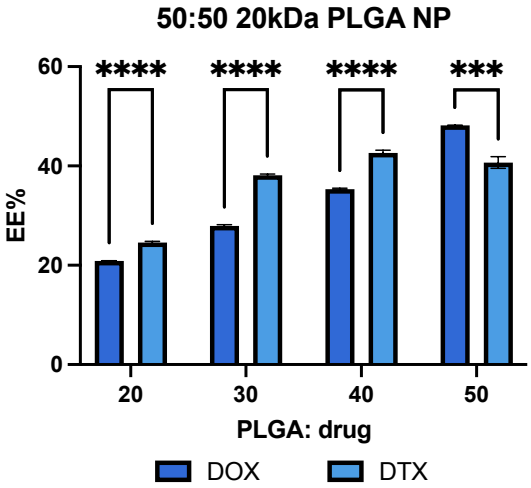
Figure 3.6. Effect of PLGA: drug ratio on DTX EE%: A) DTX EE% of 50:50 20kDa PLGA NP. B) DTX EE% of 50:50 40kDa PLGA NP. C) DTX EE% of 75:25 20kDa PLGA NP. D) DTX EE% of 75:25 40kDa PLGA NP. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001. Data represent the mean ± SD (N=3).

As shown in Figure 3.6, DTX EE% increases with PLGA concentration when PLGA MW, LA/GA ratio, and drug (logP) are fixed. The increment in EE% can be explained by the same reason as for Figure 3.4. However, DTX EE% of formulation 50_50_20k is lower than that of 40_50_20k, and DTX EE% of 50_75_20k has no significant difference from that of 40_75_20k. This could be explained by PLGA saturation in oil phase as PLGA concentration increased. The

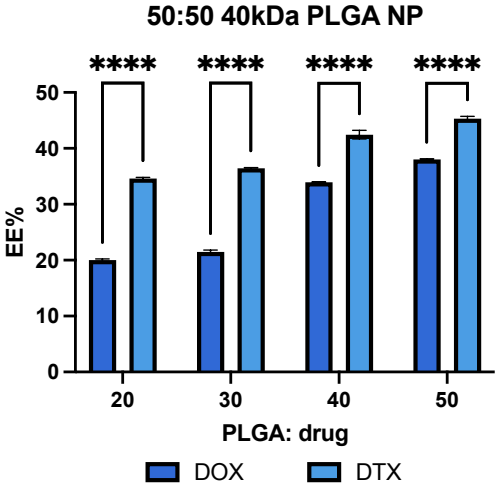
extra PLGA failed to continue forming emulsion droplets. As DTX is hydrophobic and was added to PLGA solution in the oil phase, the strong interaction between PLGA and DTX resulted in loss of DTX along with the extra PLGA, resulted in lower EE%. Additional independent experiments are necessary to confirm the variations between repeats.

3.2.1.3 DOX vs. DTX EE%

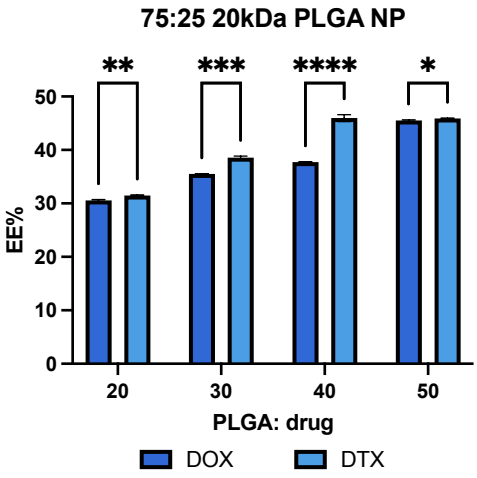
A.



B.



C.



D.

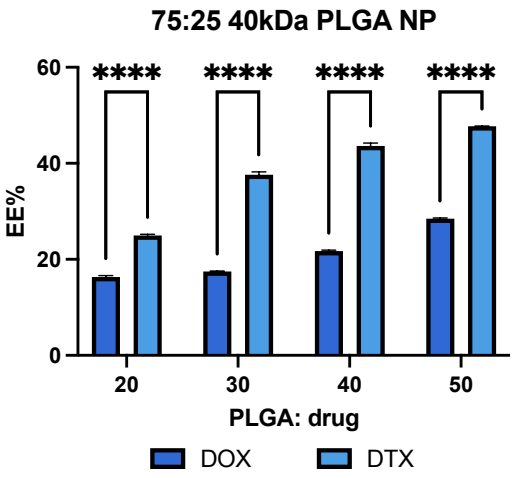


Figure 3.7. DTX and DOX EE% at different PLGA MW and LA/GA ratio. A) Drug EE% at 50:50 20kDa PLGA. B) Drug EE% at 50:50 40kDa PLGA. C) Drug EE% at 75:25 20kDa PLGA. D) Drug EE% at 75:25 40kDa PLGA. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001. Data represent the mean ± SD (N=3).

Under the same NP composition, DTX EE% are in general higher than DOX EE% due to the high logP of DTX and the solution phase DTX is in. This is because drugs with different logP have different solubilities and partitioning in water or oil phase. Hydrophilic drug easily migrates from the first water phase to the second water phase and eventually be released into the solution, while hydrophobic DTX tends to stay in the oil phase with more interactions with the hydrophobic PLGA matrix. As mentioned above, DTX EE% in 50_50_20k NP is lower than formulation 40_50_20k due to PLGA saturation. This reflects in a significantly lower (p<0.001) DTX EE% than DOX EE% in 50_50_20k NP.

3.2.2 IC50

Considering the subsequent comparisons between different NP formulations, absolute IC50 was used instead of relative IC50. In a Hill model, the top was set to 100% cell viability that represents no cell growth inhibition, and the baseline was set to 0% cell viability that represents 100% cell growth inhibition. The absolute IC50 is therefore the drug concentration that inhibits 50% cell growth. The lower and upper end of the sigmoidal curve in Figure 3.8 reach plateaus at baseline and top. The dose-response Hill model was built following Equation 9 using a nonlinear regression fit for log[absolute IC50] in GraphPad Prism 9[108].

$$\text{Cell viability} = \text{Bottom} + \frac{\text{Top} - \text{Bottom}}{1+10^{[(\log absolute\ IC50 - \log drug\ conc.)\text{Hill Slope} + \log\left(\frac{Top-Bottom}{50-Bottom}-1\right)]}} \quad (9)$$

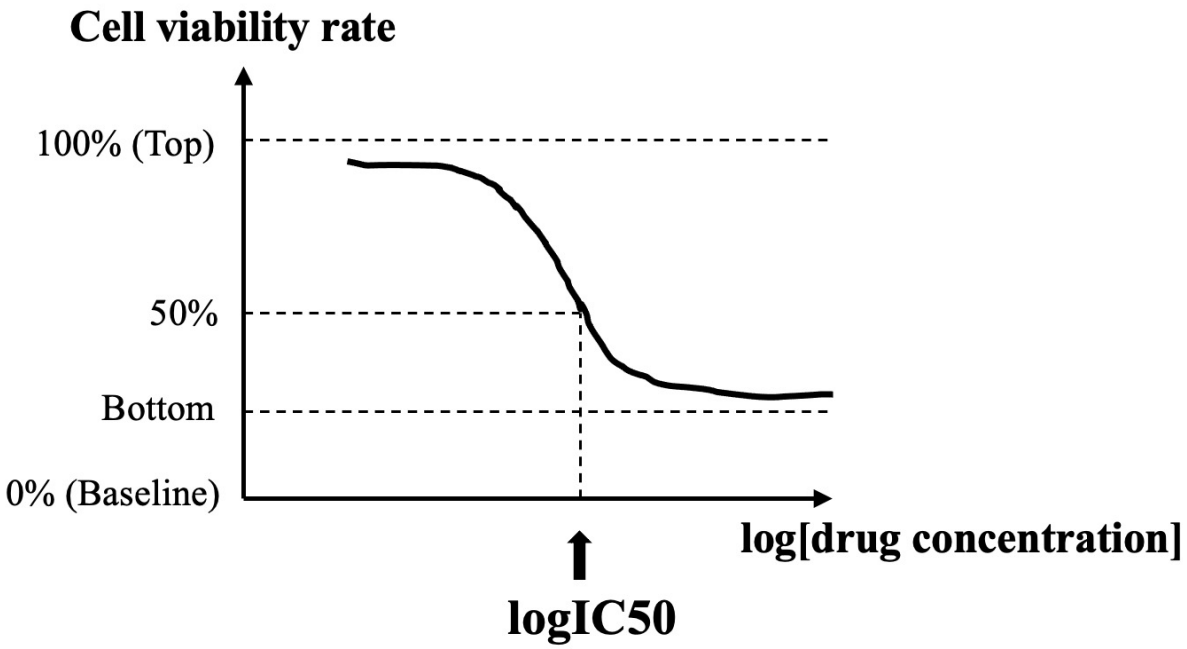


Figure 3.8. Schematic of absolute IC50.

The free drug IC50 values were obtained under the same cell viability test protocol as drug-loaded NPs. Free DOX IC50 are either lower than or not significantly different from DOX NP IC50 (Appendix C Figure C1). Free DTX IC50 are also either lower than or not significantly different from most of the DTX NP IC50, except for formulation 50_50_20k (Appendix C Figure C2). IC50 relies on how drugs were taken up by cells. The drug delivery pathways for free drug and NPs are quite different. As mentioned before, free drugs passively diffuse into the cells, while NPs with size greater than 200 nm undergo macropinocytosis or phagocytosis following entry into the digestive endosome. For the same type of drug, its cytotoxicity in different formulations also depends on how much of it accumulated in the cells within the fixed incubation time[109]. In other words, the more drug being released within cells, the greater the therapeutic effects, hence a lower IC50. The drug release rate from NPs was not determined in this study and the cell uptake profile

also remains unclear. According to the IC50 results, we could assume that the cellular drug accumulation of free drugs is higher than that of most of the drug-loaded NPs within the same amount of incubation time (48 hrs). The low drug accumulation of NPs might because of the slow drug release rate and NP decomposition rate.

A. B.

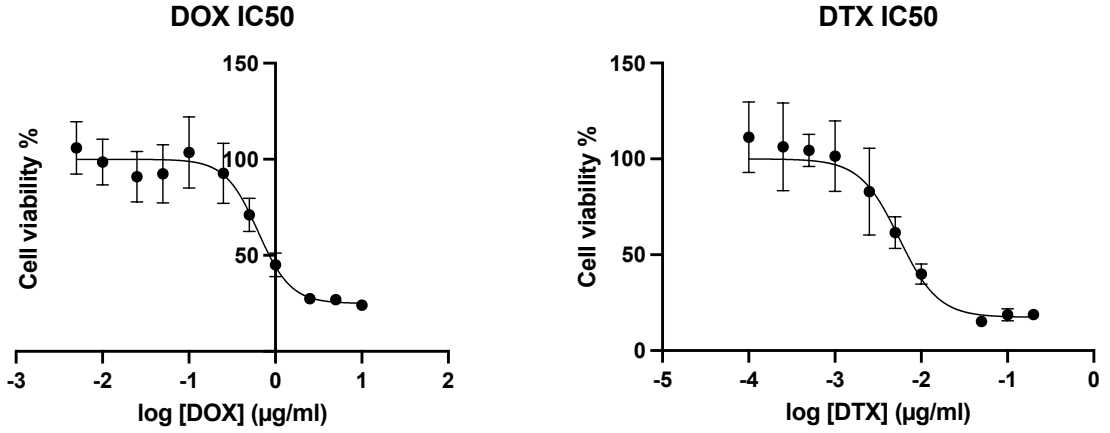
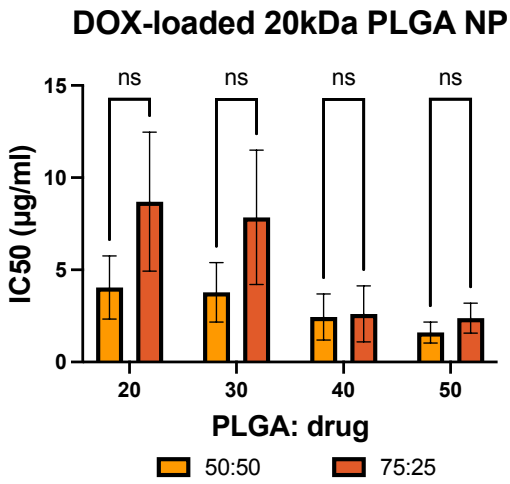


Figure 3.9. A) Dose response curve of free DOX drug. IC50 is 0.85±0.23 µg/mL. B) Dose response curve of free DTX drug. IC50 is 0.007±0.002 µg/mL. The free drug cytotoxicity tests follow the same procedure as drug-loaded NP cytotoxicity tests. Data represent the mean ± SD (N=6). Standard deviation error bars that are smaller than the data point symbol are hidden.

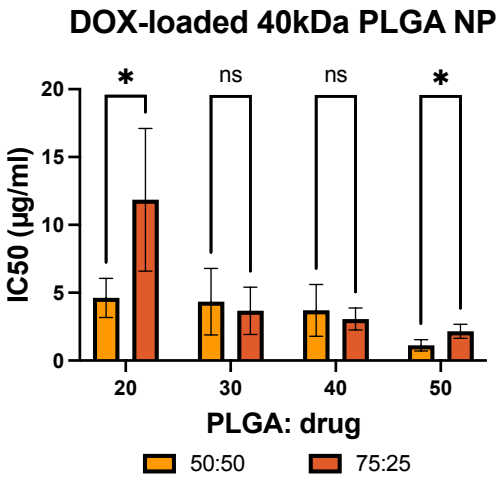
NP composition should be well-evaluated and optimized if one aims to enhance the therapeutic efficacy of drug-loaded NPs compared to free drugs, in terms of increasing cellular accumulation and controlled drug release. Release study and cellular uptake study are required upon demands. These were not investigated in this study but should be involved in the future.

3.2.2.1 DOX IC50

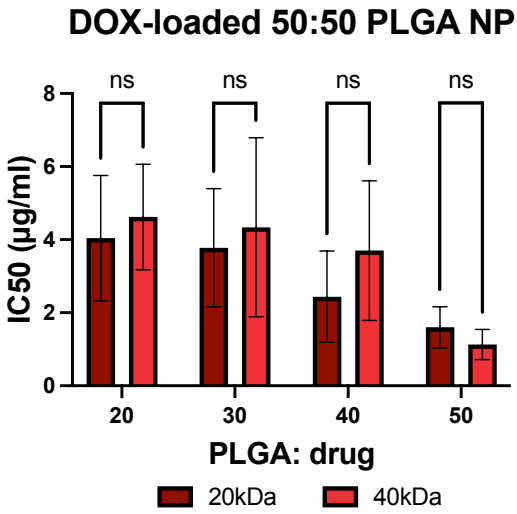
A.



B.



C.



D.

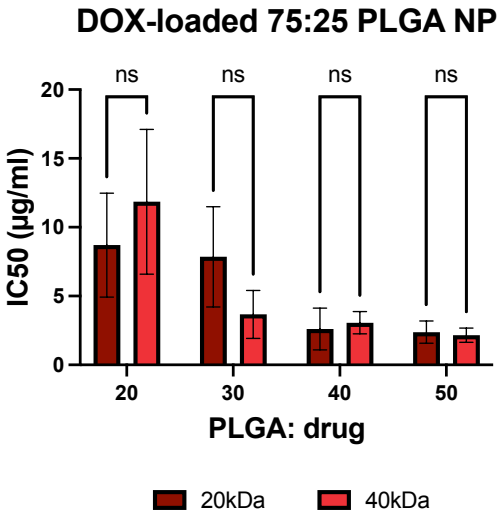
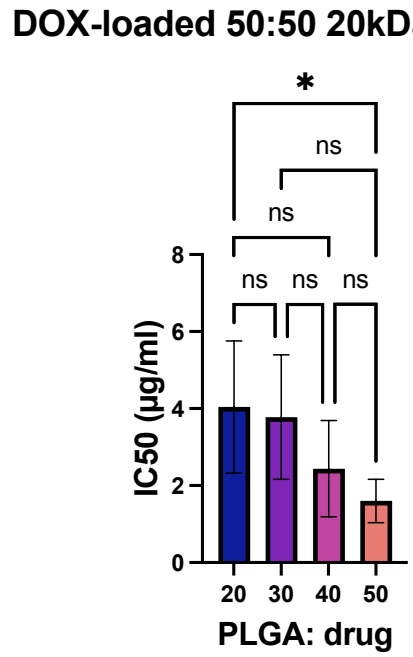
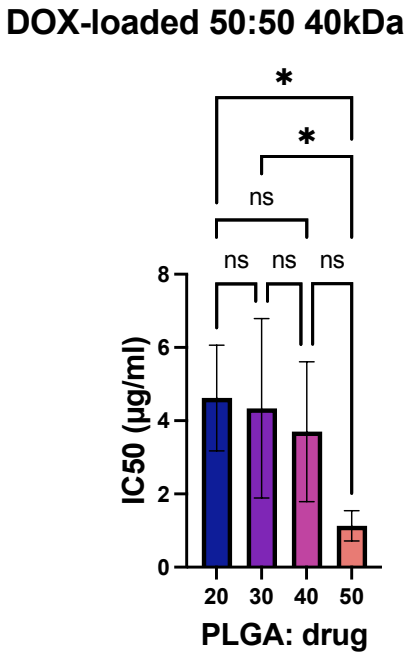


Figure 3.10. Effect of LA/GA ratio on DOX IC50: A) DOX IC50 of 20kDa PLGA NP. B) DOX IC50 of 40kDa PLGA NP. Effect of PLGA MW on DOX IC50: C) DOX IC50 of 50:50 PLGA NP. D) DOX IC50 of 75:25 PLGA NP. Multiple t test was performed. *$p<0.05$. Data represent the mean ± SD (N=6).

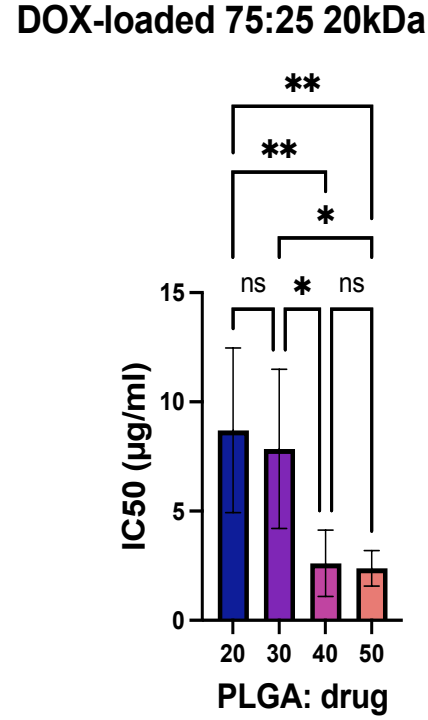
A.


DOX-loaded 50:50 20kDa PLGA NP
IC50 (µg/ml)
PLGA: drug
20
30
40
50
ns
*

B.


DOX-loaded 50:50 40kDa PLGA NP
IC50 (µg/ml)
PLGA: drug
20
30
40
50
ns
*

C.


DOX-loaded 75:25 20kDa PLGA NP
IC50 (µg/ml)
PLGA: drug
20
30
40
50
ns
*
**

D.

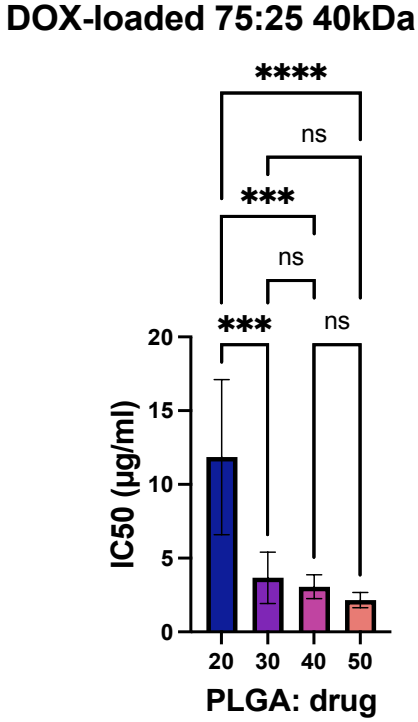

DOX-loaded 75:25 40kDa PLGA NP
IC50 (µg/ml)
PLGA: drug
20
30
40
50
ns
***
****

Figure 3.11. Effect of PLGA: drug ratio on DOX IC50: A) DOX IC50 of 50:50 20kDa PLGA NP. B) DOX IC50 of 50:50 40kDa PLGA NP. C) DOX IC50 of 75:25 20kDa PLGA NP. D) DOX IC50 of 75:25 40kDa PLGA NP. One way ANOVA test was performed. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001. Data represent the mean ± SD (N=6).

DOX is fluorescent and can be excited at 490 nm wavelength which is the same as MTS cell viability assay detection wavelength[110]. Since cells were all washed by PBS twice in both free DOX and DOX NP MTS assays, it is assumed that all residual DOX should have been washed away. Studies were performed to investigate the interference of DOX on MTS assay readings in the presence or absence of cells at A490 (Appendix B Figure B2 and B3). Results show there were no significant differences between the negative control group and up to 5μg/mL DOX in the presence of cells and 0.1 μg/mL DOX in the absence of cells.

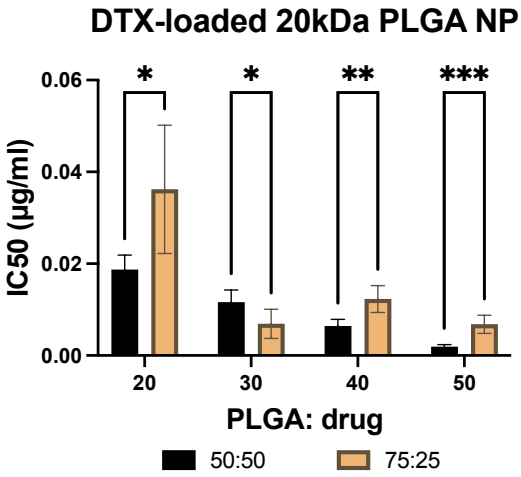
According to results in Figure 3.10, DOX NP IC50 is generally uncorrelated to PLGA LA/GA ratio (Figure 3.10A and B) or PLGA MW (Figure 3.10 C and D). It is likely that the chosen PLGA LA/GA ratio and PLGA MW do not have significant impact on the NP cellular accumulation or drug release that are related to NP IC50. And, because of the high standard deviation in each test group in cell viability tests, many of the differences between groups are statically insignificant. However, we noted that 75:25 40kDa PLGA NPs with 20:1 and 50:1 PLGA: drug ratio have higher IC50 than 50:50 40kDa PLGA NPs. Since 75:25 PLGA is more hydrophobic, it is likely hydrolyzed at a slower rate during degradation. Less drug accumulates in cells, so, more drug is needed to inhibit 50% of cell growth in 75:25 PLGA NPs.

On the other hand, DOX NP IC50 generally decreases as PLGA: drug ratio increases (Figure 3.11). Recall that the DOX NP EE% increases as PLGA: drug ratio increases (Figure 3.4).

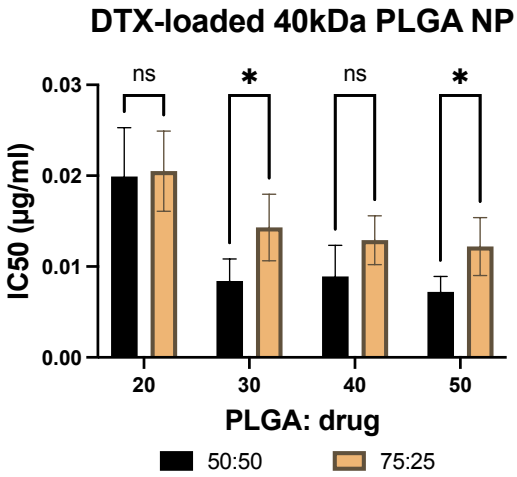
EE% also determines how much drug was delivered to cells, but not necessarily 100% of the drug is released and accumulated in cells. Combining the results in Figure 3.11, we observed that the more drug was encapsulated, the lower the NP IC50. We could assume that more drug was released and accumulated in cells at higher EE% as PLGA: drug ratio increased. Additional drug release or cellular uptake studies are necessary to support the conclusion.

3.2.2.2 DTX IC50

A.



B.



C.

D.

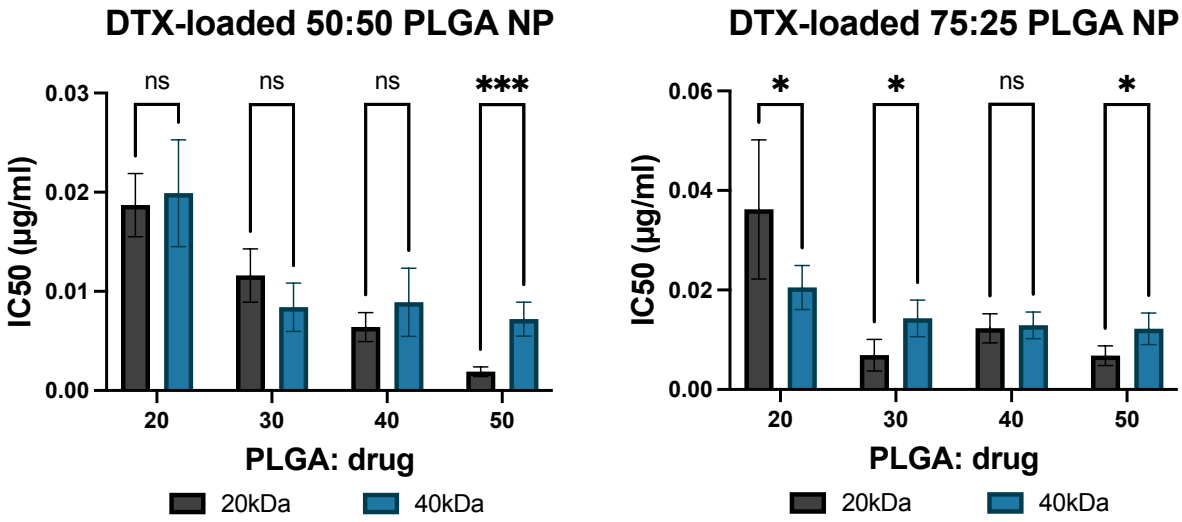
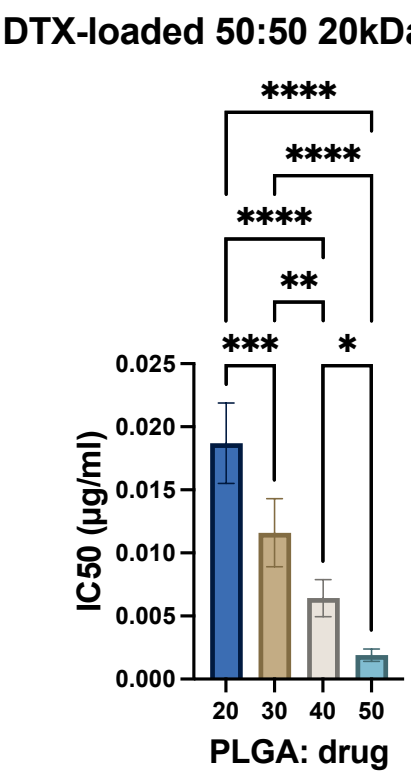
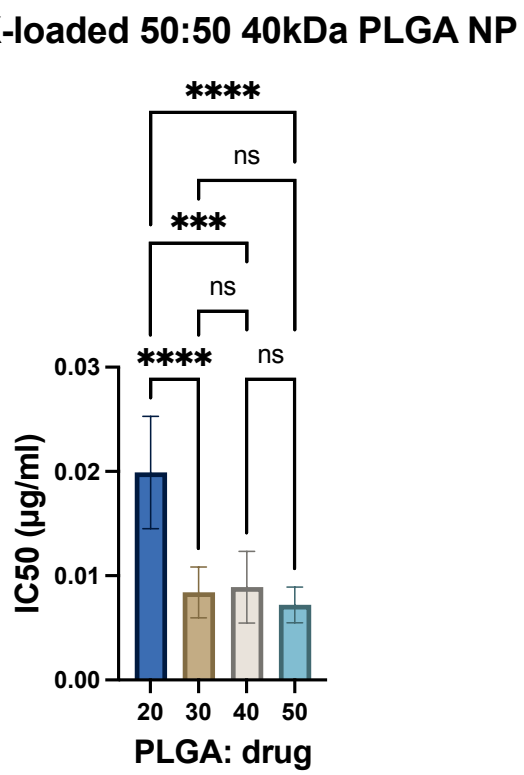


Figure 3.12. Effect of LA/GA ratio on DTX IC50: A) DTX IC50 of 20kDa PLGA NP. B) DTX IC50 of 40kDa PLGA NP. Effect of PLGA MW on DTX IC50: C) DTX IC50 of 50:50 PLGA NP. D) DTX IC50 of 75:25 PLGA NP. Multiple t test was performed. *p<0.05, **p<0.01, ***p<0.001. Data represent the mean ± SD (N=6).

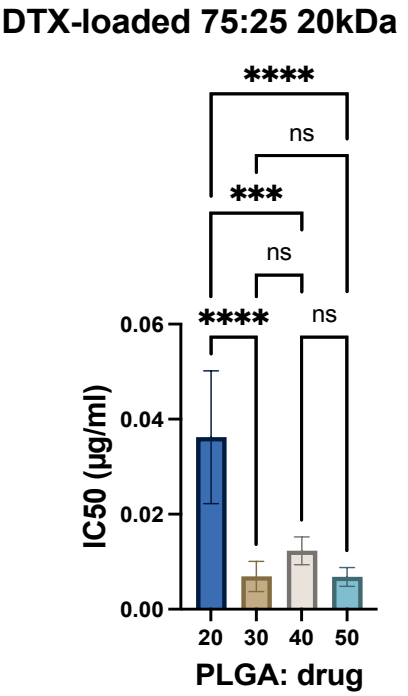
A. B.


DTX-loaded 50:50 20kDa PLGA NP
****
****
****
**
***
*
0.025
0.020
0.015
0.010
0.005
0.000
IC50 (µg/ml)
20
30
40
50
PLGA: drug


DTX-loaded 50:50 40kDa PLGA NP
****
ns
***
ns
****
ns
0.03
0.02
0.01
0.00
IC50 (µg/ml)
20
30
40
50
PLGA: drug

C.

D.


DTX-loaded 75:25 20kDa PLGA NP
****
ns
***
ns
****
ns
0.06
0.04
0.02
0.00
IC50 (µg/ml)
20
30
40
50
PLGA: drug

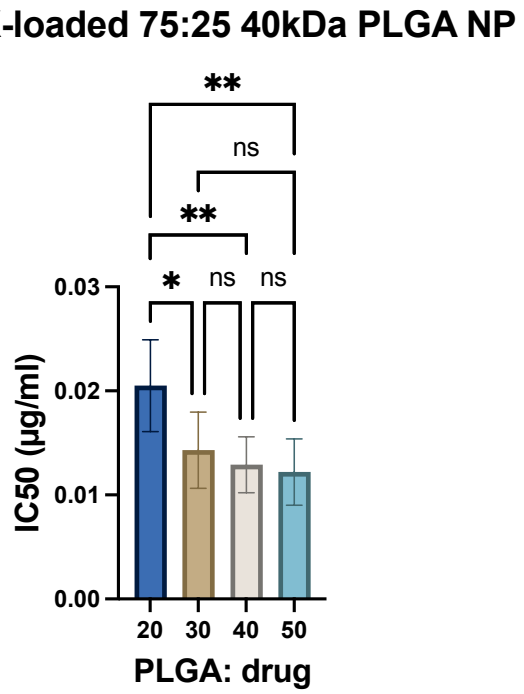

DTX-loaded 75:25 40kDa PLGA NP
**
ns
**
*
ns
ns
0.03
0.02
0.01
0.00
IC50 (µg/ml)
20
30
40
50
PLGA: drug

Figure 3.13. Effect of PLGA: drug ratio on DTX IC50: A) DTX IC50 of 50:50 20kDa PLGA NP. B) DTX IC50 of 50:50 40kDa PLGA NP. C) DTX IC50 of 75:25 20kDa PLGA NP. D) DTX IC50 of 75:25 40kDa PLGA NP. One way ANOVA test was performed. *$p<0.05$, **$p<0.01$, ***$p<0.001$, ****$p<0.0001$. Data represent the mean ± SD (N=6).

DTX NP IC50 also does not have general association with PLGA LA/GA ratio or PLGA MW (Figure 3.12). In Figure 3.12A, the 75:25 20kDa PLGA NPs have higher IC50 than 50:50 PLGA NPs, except for formulations at 30:1 PLGA: drug ratio. The higher IC50 might be due to slower drug release. 75:25 PLGA tends to degrade slower because of its hydrophobic nature, thus, results in lower drug accumulation and higher IC50. For the same reason, 75:25 40kDa PLGA NPs at 30:1 and 50:1 PLGA: drug ratio have higher IC50 than 50:50 40kDa PLGA NPs in Figure 3.12B as well. NP degradation rate is reduced as PLGA MW increases because longer polymer chain requires more energy to be hydrolyzed. Higher drug concentration is required to achieve the same cytotoxicity. This explains the higher IC50 in 50:50 40kDa PLGA NP than 20kDa PLGA NP at 50:1 PLGA: drug ratio in Figure 3.12C. However, there is no relationship between DTX NP IC50 and PLGA MW in 75:25 PLGA NPs in Figure 3.12D. As shown in Figure 3.13, DTX NP IC50 generally decreases as PLGA: drug ratio increases, with high statistical significances ($p<0.05$), while a few other data have no statistical significances between each other. This is similar to what was observed in Figure 3.11 for DOX IC50. As higher EE% was achieved in high PLGA: drug ratio NP, more DTX was delivered to cells. There is higher chance that more drug was released and accumulated as PLGA: drug ratio increases. Since IC50 can be associated with drug accumulation, a higher amount of drug being delivered to cells may cause higher therapeutic effect that leads to a lower IC50.

3.3 GP model

3.3.1 Data preprocessing and split

As prior mean was set to 0 in GP models, all data have also been standardized to a mean of 0 and standard deviation of 1 to ease the computation and to accelerate the model training[111]. Considering the different mechanism of action DOX and DTX inherit, the IC50 models were built independently for each drug. Data were split into training and testing set at a ratio of 8:2 in each of the GP models, that are 25 training points and 7 testing points in EE% model, and 12 training points and 4 testing points in each IC50 model. Training set was used to update GP posterior distribution, and testing set was used to evaluate model prediction performance.

3.3.2 Kernel functions

To make sure that physicochemical properties (x variables) are computer readable data, PLGA: drug ratios were converted to numerical values 20, 30, 40, and 50. And LA/GA ratio were converted to 1 and 3.

Kernel function in GaussianProgressRegressor class in scikit-learn structures the true function by specifying the covariance between two input data. The hyperparameters length scale and nu that control the kernel function smoothness were optimized to better fit given training data. Upon data training, the posterior kernel estimates the prediction uncertainty (indicated by confidence interval) of the output values[112]. The uncertainty in the region that contains more observed data is likely to be lower than that in the region with less observed data.

Compared to non-stationary kernel, stationary kernel simply finds the similarities between two data points from their distances rather than their values or the exact locations in input space[91,92].

We tested both Matérn kernel and RBF kernel, where RBF is a special case of Matérn when the hyperparameter υ approaches infinity. The stationary Matérn kernel outcompeted RBF kernel with higher prediction accuracy with respect to higher R^2 in predicted-y versus observed y linear regression plot in each GP model. The hyperparameters 'length_scale', 'length_scale_bounds', and 'nu' are listed in Table 2.1 in Section 2.7. The length scale controls the smoothness of the kernel function. A smaller length scale provides a function that loses the correlations between y values at a shorter distance between two x values. 'nu' controls the function differentiability that also accounts for function smoothness and its ability of taking derivatives at a point. 'nu' was set to 0.5 for all models, with a kernel of absolute exponential function. The initial length scale was set as 1 in each model. With the goal of maximizing the log-marginal likelihood function (or minimizing negative log-marginal likelihood function), the hyperparameter optimization restarted 20 times following the setup of 'n_restarts_optimizer' within the 'length_scale_bounds'.

For better estimation of the true function, the addition of data noise enables more accurate modeling. A White Kernel function of y variable noise ε can be added to the kernel matrix diagonal. The noise should be calculated as the average variance of the output values. It can be the measurement noises or the noises between repeated experiments. This method is worth trying in the future to obtain a better understanding of the data.

3.3.3 EE% prediction

3.3.3.1 Predicted EE% in testing set

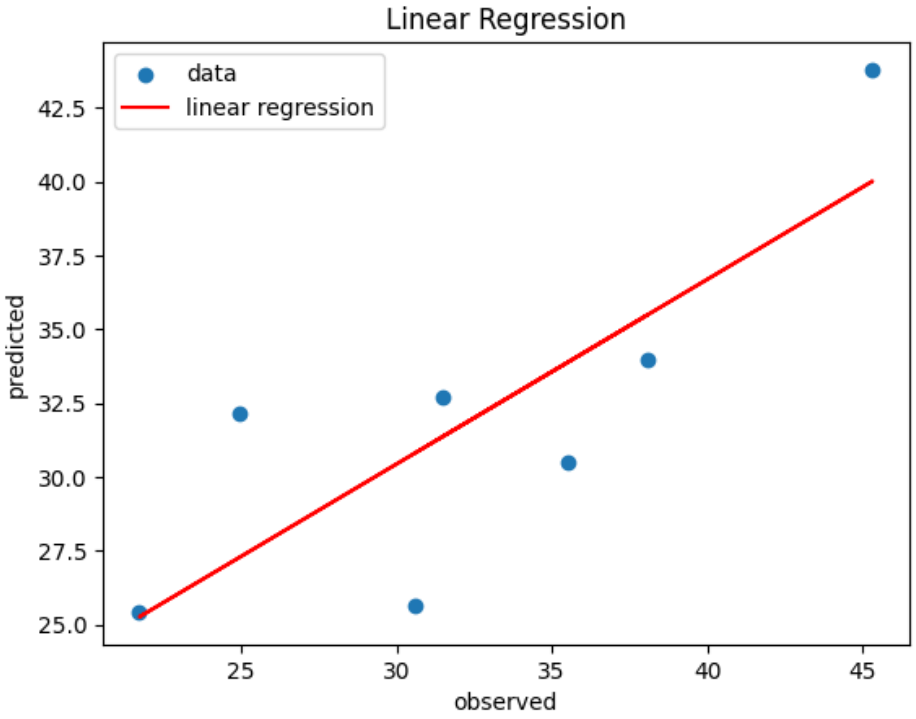


Figure 3.14. Regression curve of predicted EE% versus observed EE% in testing set. Regression line function: $y = 0.63x+11.68$ ($R^2= 0.65$).

Table 3.2. Observed and predicted EE% in testing set. Bias is calculated as the absolute value of 100%*(Predicted EE% – Observed EE%)/Observed EE%.

	Observed EE%	**Predicted EE%**	**Bias%**
DOX 20_75_20k	30.58	25.63	16.19
DOX 30_75_20k	35.53	30.52	14.10
DOX 40_75_40k	21.73	25.42	16.98

DTX 20_75_20k	31.49	32.68	3.78
DTX 20_75_40k	24.96	32.13	28.73
DTX 30_50_20k	38.10	33.97	10.84
DTX 50_50_40k	45.32	43.79	3.38

Table 3.3. RMSE value for both training and testing models in EE% prediction.

Model	**RMSE**	**Normalized RMSE**
Training	$1.49*10^{-9}$	$4.68*10^{-11}$
Testing	4.40	0.187

Table 3.4. n_restarts_optimizer value and optimized hyperparameters in EE% model.

n_restarts_optimizer	**length_scale**	**nu**
20	4.02	0.5

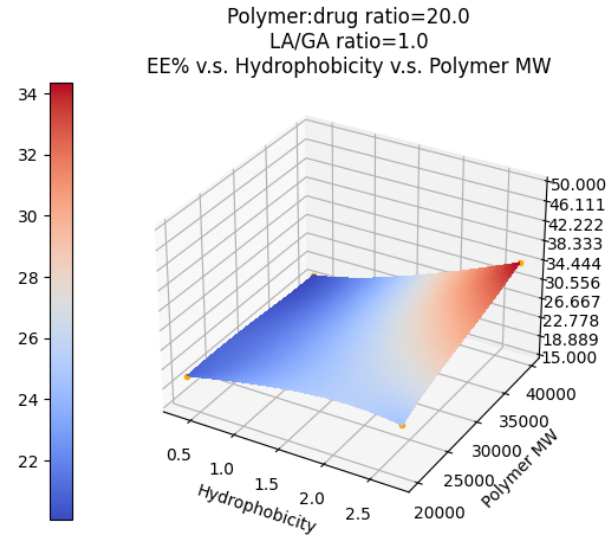
The R^2 coefficient of determination in Figure 3.14 measures how much of the variance in predicted y can be explained by the observed y. $R^2 = 0.65$ indicates a relatively good linear regression model explanation for predicted and observed EE%. However, R^2 is not a convincing metric to evaluate model prediction accuracy. The ideal linear regression line that indicates high prediction accuracy should have a slope of 1 and y-intercept close to 0. However, a high R^2 does not refer to high prediction accuracy. The biases calculated in Table 3.2 elaborates the percentage of predicted EE% that deviates from observed EE%. Among all test data, DTX 20_75_40k NP has

the largest bias at 28.73% between its observed and predicted EE%. Additionally, we should be focusing more on the root mean square error (RMSE) to explain model performance (Equation 10). It calculates the standard deviation of the prediction error between predicted and observed y values. y_i is the observed y value, $\widehat{y_i}$ is the predicted y value, and N is the number of data points. The lower RMSE, the higher prediction accuracy of the model. To compare RMSE between models, it is normalized to the range of observed y values in each model. Normalized-RMSE less than 1 is acceptable, meaning the prediction error is smaller than the observed output range. However normalized-RMSE larger than 1 indicates poor prediction. Training set normalized-RMSE is much lower than that in testing set due to the high flexibility of GP in learning data. The normalized-RMSE of EE% testing model is 0.187 (Table 3.3). Liemohn et al. stated that the model prediction accuracy is good when RMSE is lower than the standard deviations of observed values and the predicted values[113]. However, they did not specify a value or relative percentage to evaluate the acceptability of normalized RMSE. The final hyperparameter length scale for kernel is 4.02 in EE% model (Table 3.4).

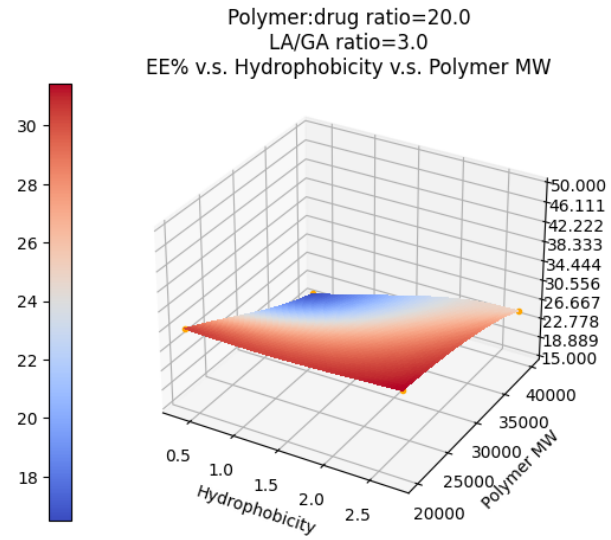
$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \widehat{y_i})^2}{N}} \tag{10}$$

3.3.3.2 EE% 3D surface response curve

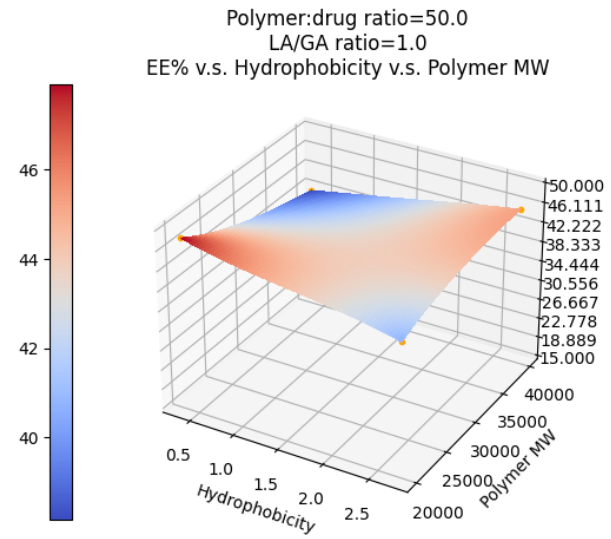
A. B.


Polymer:drug ratio=20.0
LA/GA ratio=1.0
EE% v.s. Hydrophobicity v.s. Polymer MW
34
32
30
28
26
24
22
50.000
46.111
42.222
38.333
34.444
30.556
26.667
22.778
18.889
15.000
40000
35000
30000
25000
20000
Polymer MW
0.5
1.0
1.5
2.0
2.5
Hydrophobicity

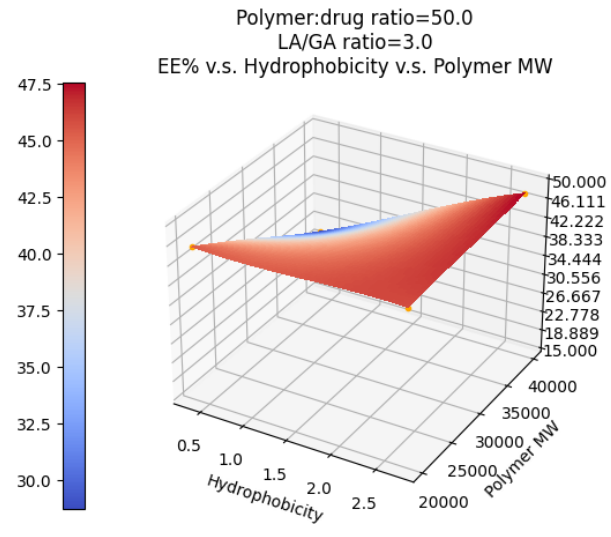
C.


Polymer:drug ratio=20.0
LA/GA ratio=3.0
EE% v.s. Hydrophobicity v.s. Polymer MW
30
28
26
24
22
20
18
50.000
46.111
42.222
38.333
34.444
30.556
26.667
22.778
18.889
15.000
40000
35000
30000
25000
20000
Polymer MW
0.5
1.0
1.5
2.0
2.5
Hydrophobicity

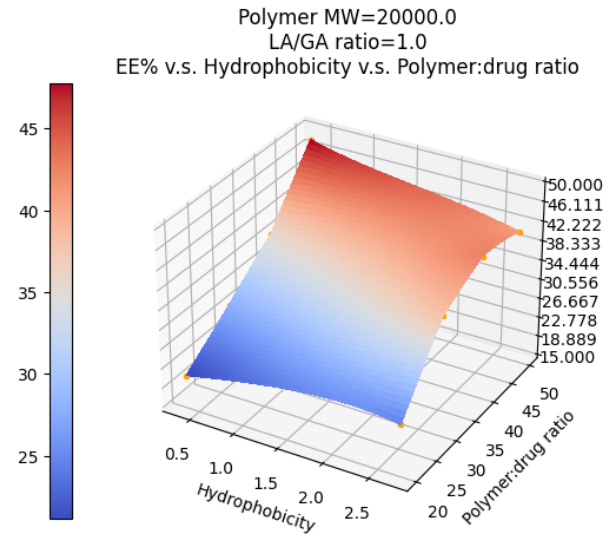
D.


Polymer:drug ratio=50.0
LA/GA ratio=1.0
EE% v.s. Hydrophobicity v.s. Polymer MW
46
44
42
40
50.000
46.111
42.222
38.333
34.444
30.556
26.667
22.778
18.889
15.000
40000
35000
30000
25000
20000
Polymer MW
0.5
1.0
1.5
2.0
2.5
Hydrophobicity

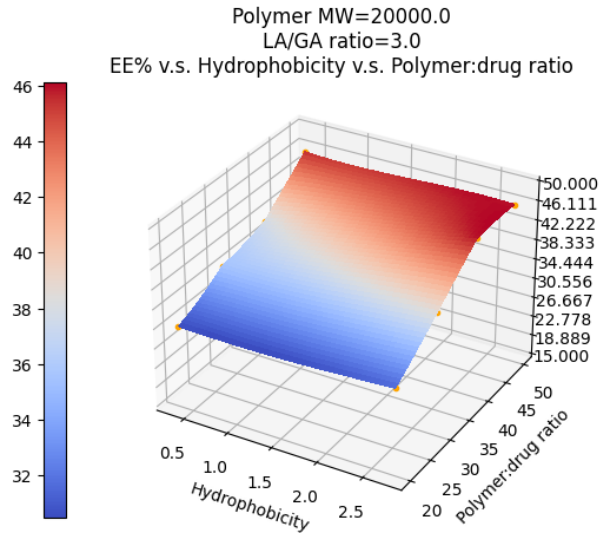
E.


Polymer:drug ratio=50.0
LA/GA ratio=3.0
EE% v.s. Hydrophobicity v.s. Polymer MW
47.5
45.0
42.5
40.0
37.5
35.0
32.5
30.0
50.000
46.111
42.222
38.333
34.444
30.556
26.667
22.778
18.889
15.000
40000
35000
30000
25000
20000
Polymer MW
0.5
1.0
1.5
2.0
2.5
Hydrophobicity

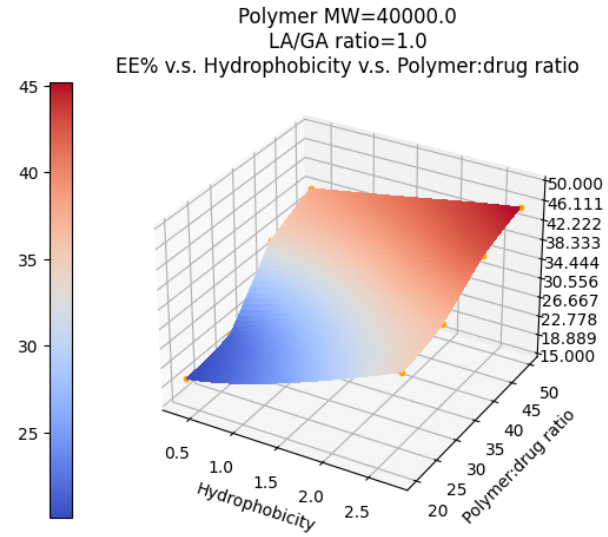
F.


Polymer MW=20000.0
LA/GA ratio=1.0
EE% v.s. Hydrophobicity v.s. Polymer:drug ratio
Hydrophobicity
Polymer:drug ratio

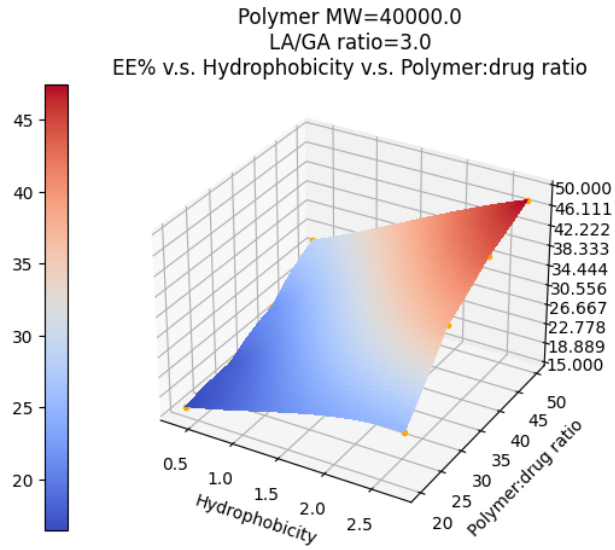
G.


Polymer MW=20000.0
LA/GA ratio=3.0
EE% v.s. Hydrophobicity v.s. Polymer:drug ratio
Hydrophobicity
Polymer:drug ratio

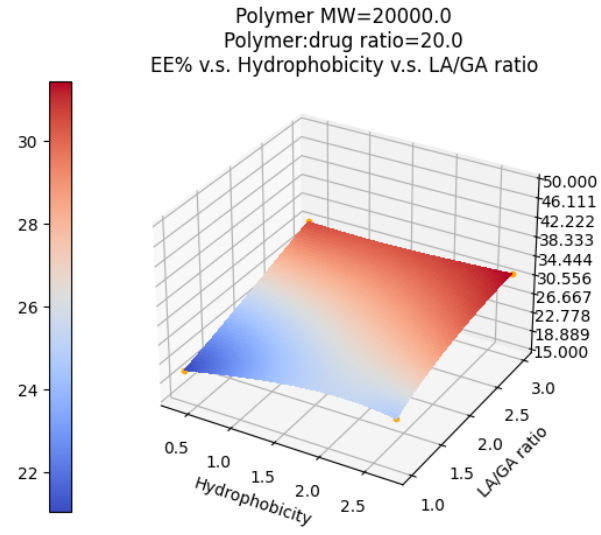
H.


Polymer MW=40000.0
LA/GA ratio=1.0
EE% v.s. Hydrophobicity v.s. Polymer:drug ratio
Hydrophobicity
Polymer:drug ratio

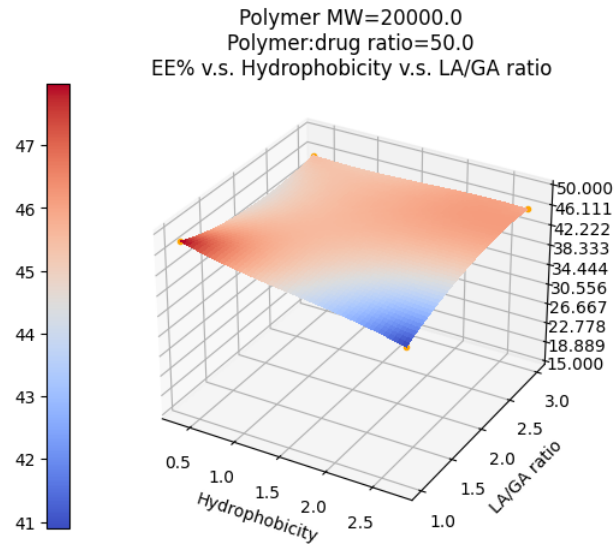
I.


Polymer MW=40000.0
LA/GA ratio=3.0
EE% v.s. Hydrophobicity v.s. Polymer:drug ratio
Hydrophobicity
Polymer:drug ratio

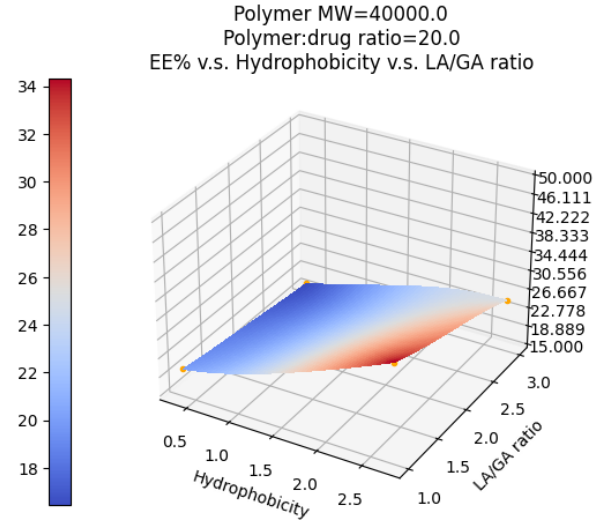
J.


Polymer MW=20000.0
Polymer:drug ratio=20.0
EE% v.s. Hydrophobicity v.s. LA/GA ratio
Hydrophobicity
LA/GA ratio

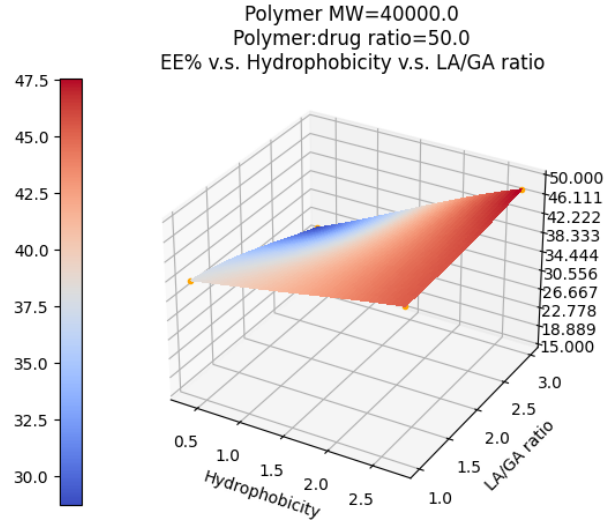
K.


Polymer MW=20000.0
Polymer:drug ratio=50.0
EE% v.s. Hydrophobicity v.s. LA/GA ratio
Hydrophobicity
LA/GA ratio

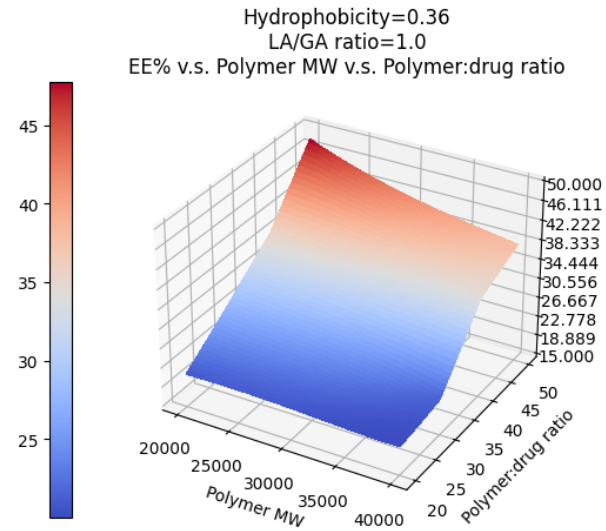
L.


Polymer MW=40000.0
Polymer:drug ratio=20.0
EE% v.s. Hydrophobicity v.s. LA/GA ratio
Hydrophobicity
LA/GA ratio

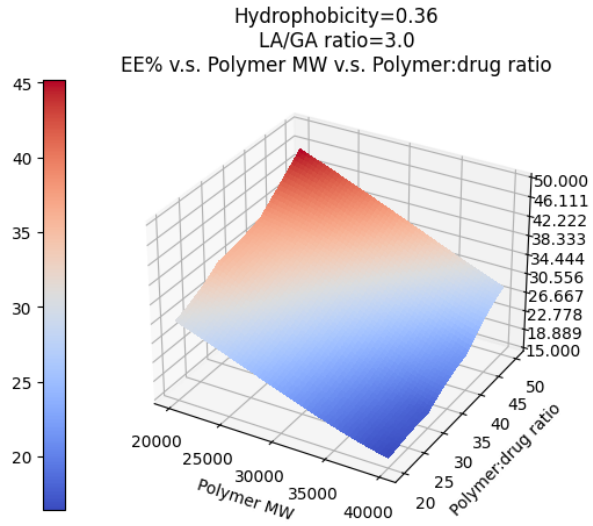
M.


Polymer MW=40000.0
Polymer:drug ratio=50.0
EE% v.s. Hydrophobicity v.s. LA/GA ratio
Hydrophobicity
LA/GA ratio

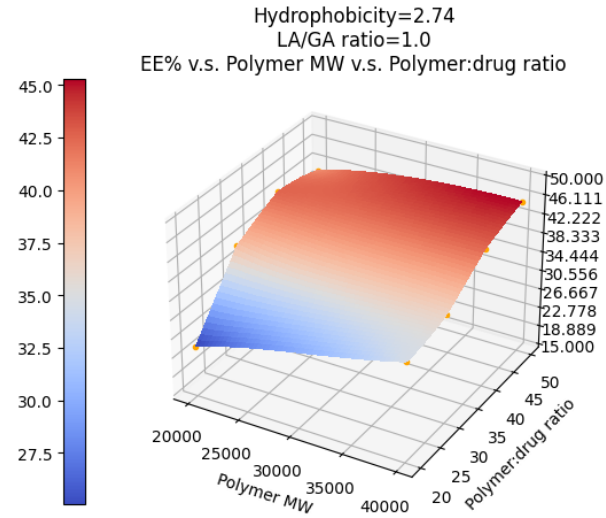
N.

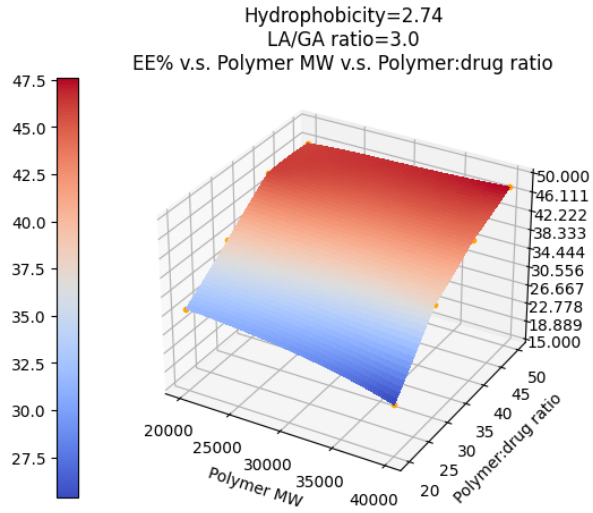

Hydrophobicity=0.36
LA/GA ratio=1.0
EE% v.s. Polymer MW v.s. Polymer:drug ratio
Polymer MW
Polymer:drug ratio

O.

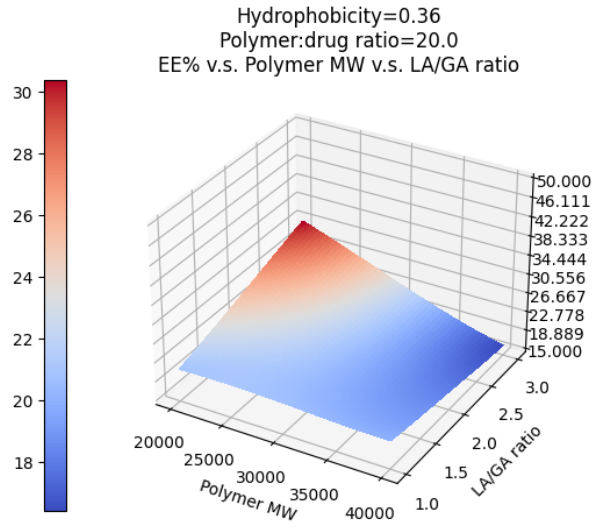

Hydrophobicity=0.36
LA/GA ratio=3.0
EE% v.s. Polymer MW v.s. Polymer:drug ratio
Polymer MW
Polymer:drug ratio

P.

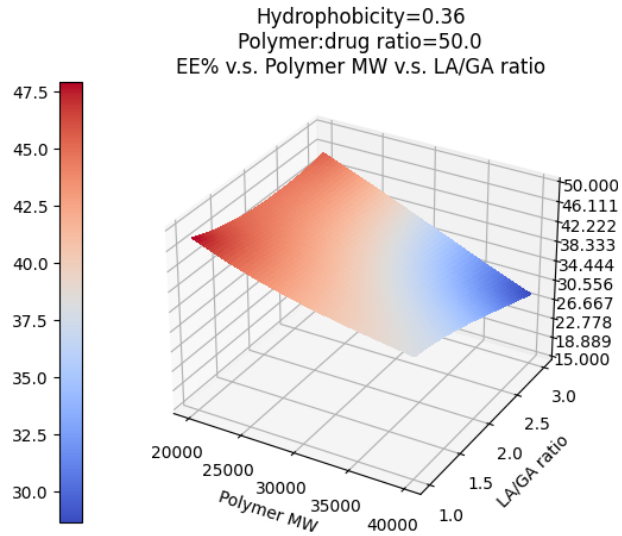

Hydrophobicity=2.74
LA/GA ratio=1.0
EE% v.s. Polymer MW v.s. Polymer:drug ratio
Polymer MW
Polymer:drug ratio


Hydrophobicity=2.74
LA/GA ratio=3.0
EE% v.s. Polymer MW v.s. Polymer:drug ratio
Polymer MW
Polymer:drug ratio

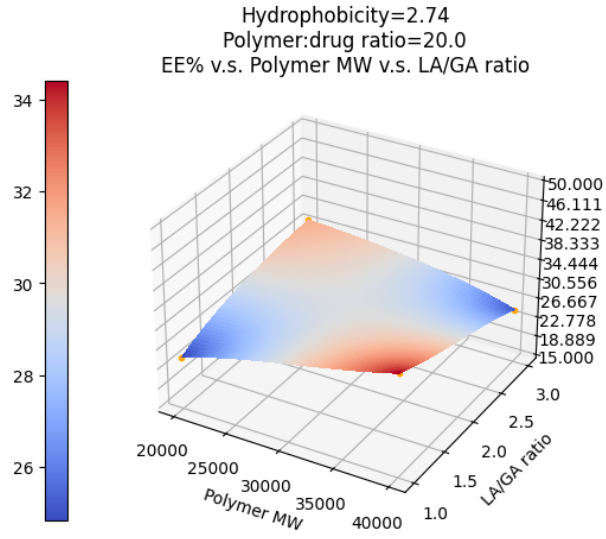
Q.


Hydrophobicity=0.36
Polymer:drug ratio=20.0
EE% v.s. Polymer MW v.s. LA/GA ratio
Polymer MW
LA/GA ratio

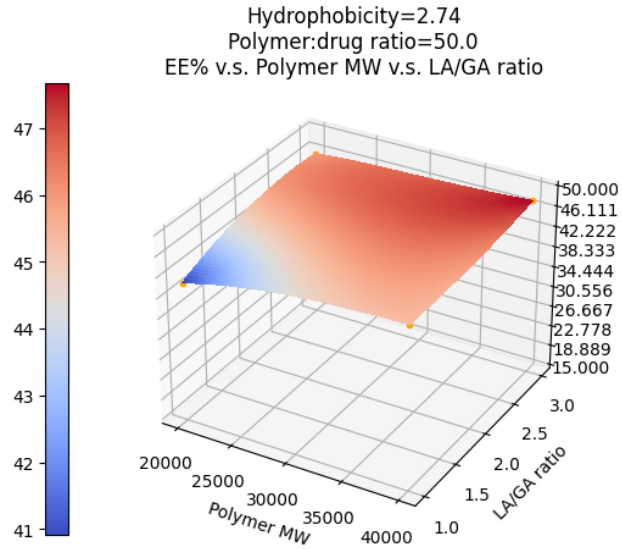
R.


Hydrophobicity=0.36
Polymer:drug ratio=50.0
EE% v.s. Polymer MW v.s. LA/GA ratio
Polymer MW
LA/GA ratio

S.

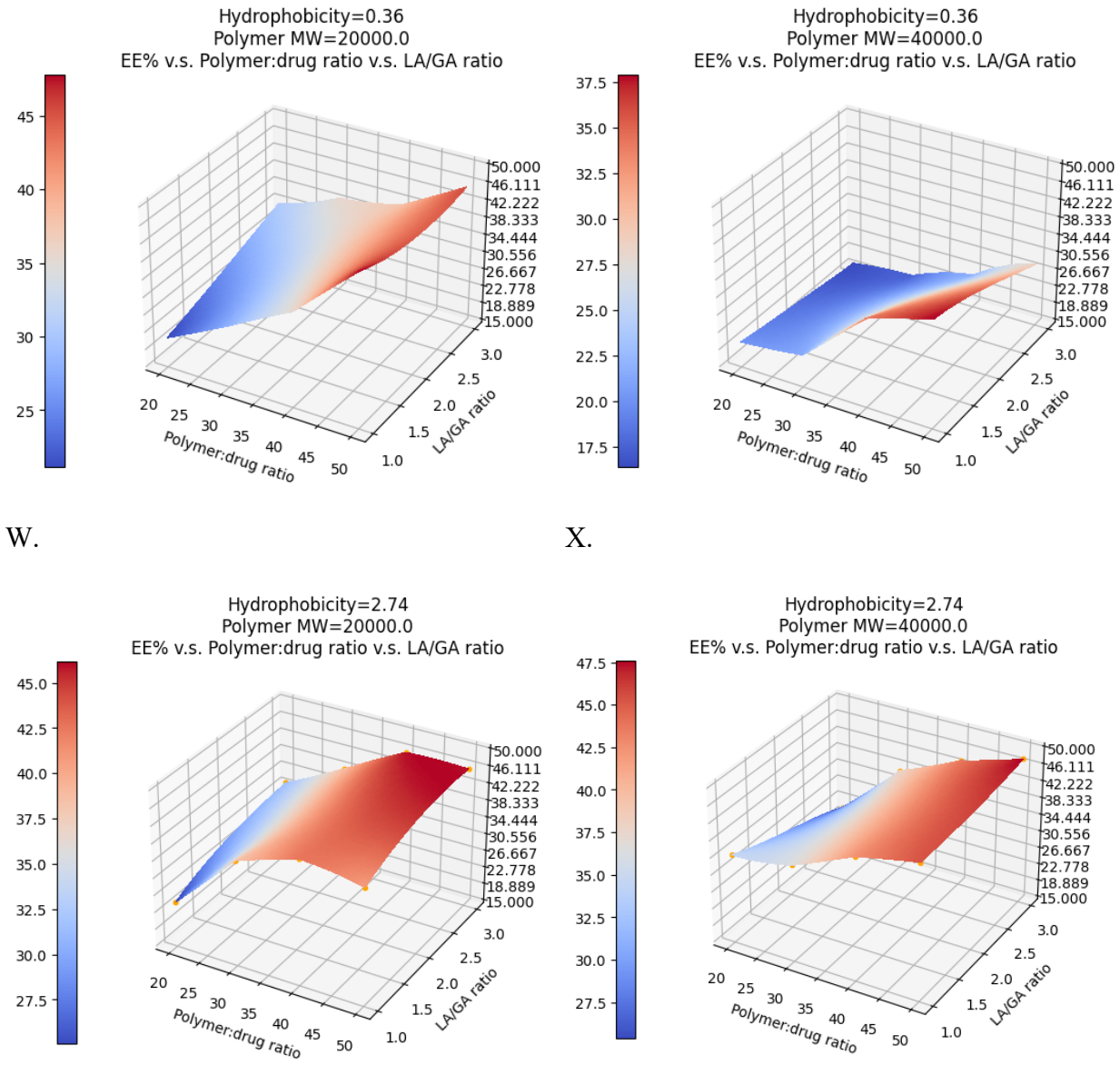

Hydrophobicity=2.74
Polymer:drug ratio=20.0
EE% v.s. Polymer MW v.s. LA/GA ratio
Polymer MW
LA/GA ratio

T.


Hydrophobicity=2.74
Polymer:drug ratio=50.0
EE% v.s. Polymer MW v.s. LA/GA ratio
Polymer MW
LA/GA ratio

U.

V.



W. X.

Figure 3.15. 3D surface response curve for NP EE% at two fixed levels of drug hydrophobicity (logP), LA/GA ratio, polymer: drug ratio, and polymer MW. The fixed levels are the lower and upper limit for each input variable.

3D surface response curves visualize the moving trend of responses under different predictor variables. Due to limited training data, all data were used for generating the curve in Figure 3.15 to better observed the trends between predictor and response variables (orange dots represent the data). Blue color in the gradient bar represents lower EE%, and red color represents

higher EE%. The surfaces in Figure 3.15 represent the predictions of EE% when two of the input variables are fixed at a lower and upper level of the domains at a time, while the other two variables lie on the x_1 and x_2 axes. The fixed levels are 0.36 and 2.74 for drug logP, 20000 and 40000 for PLGA MW, 20 and 50 for PLGA: drug ratio, and 1 and 3 for LA/GA ratio. The surfaces show the changes of EE% upon altering two of the x variables. We found that in Figure 3.15E-H, M-P, and U-X, when PLGA: drug ratio serves as one of the x axes, EE% of both drugs mostly change in higher range with PLGA: drug ratio compared to the change with the other two variables (e.g., PLGA MW or LA/GA) on the other x axis. This aligns with results in Figure 3.4 and 3.6. Other obvious trends include: in Figure 3.15D, EE% of 75:25 NP at 50:1 PLGA: drug ratio decreases greatly by over 15% towards the lowest drug logP and the highest PLGA MW. In Figure 3.15K, EE% of 40kDa PLGA NP at 20:1 PLGA: drug ratio increases by over 15% towards higher drug logP and lower LA/GA ratio. In Figure 3.15L, EE% of 40kDa NP at 50:1 PLGA: drug ratio decreases roughly by 15% towards the lowest drug logP and the highest LA/GA ratio. And in Figure 3.15R, when PLGA: drug ratio was fixed at 50:1, DOX EE% increases by around 10% as PLGA MW decreases from 40kDa to 20kDa, but varies only slightly with LA/GA ratio.

3.3.4 IC50 prediction

3.3.4.1 Predicted IC50 in testing set

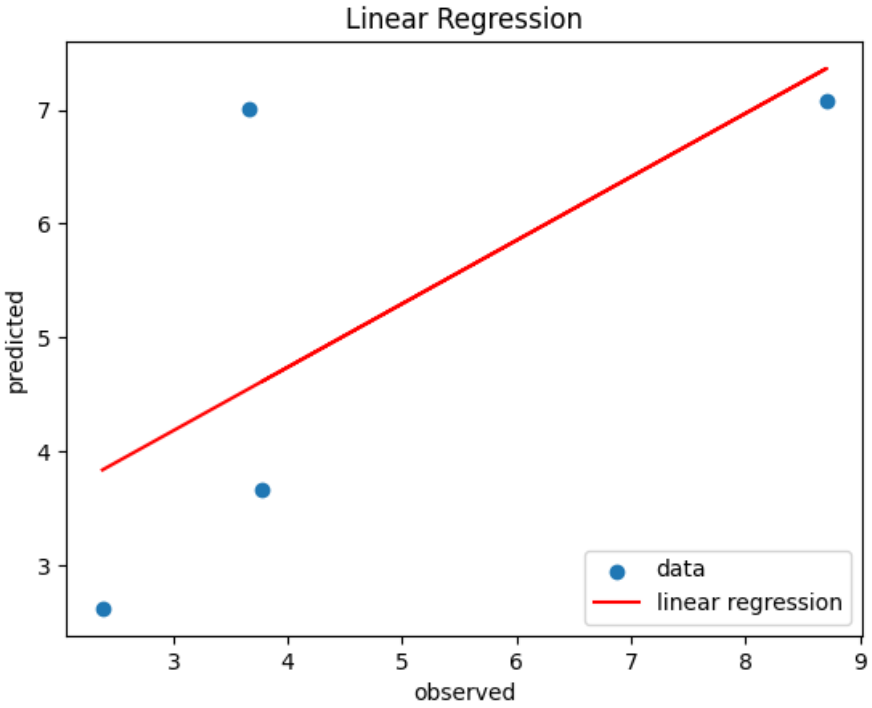


Figure 3.16. Regression curve of predicted DOX NP IC50 versus observed DOX NP IC50 in testing set. Regression line function: $y = 0.557x+2.512$ ($R^2 = 0.40$).

Table 3.5. Observed and predicted DOX NP IC50 in testing set. Bias is calculated as the absolute value of 100%*(Predicted IC50 – Observed IC50)/Observed IC50.

	Observed IC50 (μg/mL)	**Predicted IC50 (μg/mL)**	**Bias%**
DOX 20_75_20k	8.70	7.07	18.74
DOX 30_50_20k	3.78	3.66	3.17

DOX 30_75_40k	3.67	7.01	91
DOX 50_75_20k	2.38	2.62	10.08

Table 3.6. RMSE value for both training and testing models in DOX NP IC50 prediction.

Model	RMSE	Normalized RMSE
Training	$3.57*10^{-10}$	$3.33*10^{-11}$
Testing	1.87	0.296

Table 3.7. n_restarts_optimizer value and optimized hyperparameters in DOX NP IC50 model.

n_restarts_optimizer	length_scale	nu
20	1.54	0.5

The R^2 of 0.40 in DOX NP IC50 predicted vs. observed output plot (Figure 3.16) is less than 0.5, indicating a weak relationship between predicted IC50 and observed IC50. The bias between predicted and observed IC50 of DOX 30_75_40k NP is the highest among all other test data, at 91% (Table 3.5), where the predicted IC50 is almost twice of the observed IC50. This large bias might be due to the very limited knowledge of function in that region of true function. This large deviation led to the low R^2 in Figure 3.16. The normalized RMSE for DOX NP IC50 testing set in Table 3.6 is 0.296, which is higher than that of 0.187 in the EE% model. The lower normalized RMSE in EE% model could be due to the larger training set compared to the other two models, so that it generalizes testing data better. The final kernel length scale is 1.54 in this model (Table 3.7) to fit the training data.

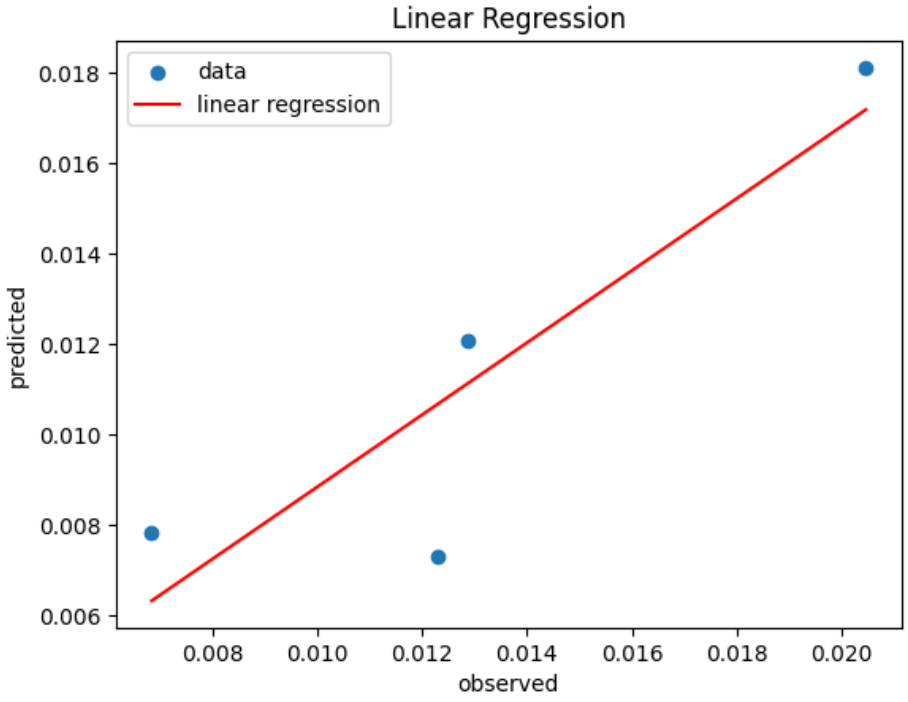


Figure 3.17. Regression curve of predicted DTX NP IC50 versus observed DTX NP IC50 in testing set. Regression line function: $y = 0.798x+0.00085$ ($R^2 = 0.66$).

Table 3.8. Observed and predicted DTX NP IC50 in testing set. Bias is calculated as the absolute value of 100%*(Predicted IC50 – Observed IC50)/Observed IC50.

	Observed IC50 (μg/mL)	**Predicted IC50 (μg/mL)**	**Bias%**
DTX 20_75_40k	0.0205	0.0181	11.71
DTX 40_75_20k	0.0123	0.0073	40.65
DTX 40_75_40k	0.0129	0.0121	6.20
DTX 50_75_20k	0.0068	0.0078	14.71

Table 3.9. RMSE value for both training and testing models in DTX NP IC50 prediction.

Model	RMSE	Normalized RMSE
Training	$1.74*10^{-12}$	$5.08*10^{-11}$
Testing	0.0028	0.206

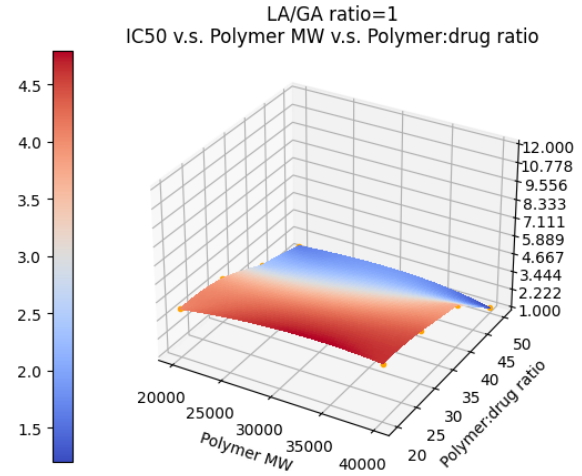
Table 3.10. n_restarts_optimizer value and optimized hyperparameters in DTX NP IC50 model.

n_restarts_optimizer	length_scale	nu
20	2	0.5

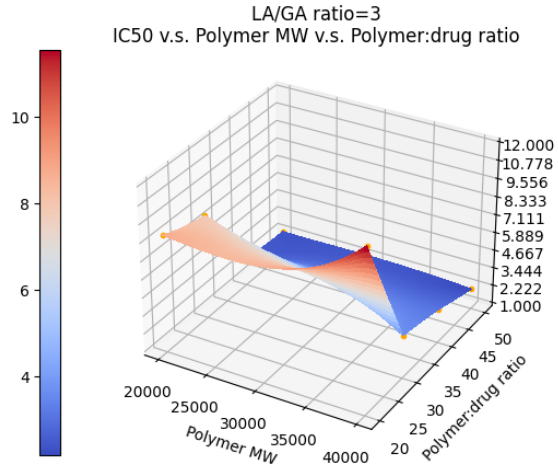
In DTX NP IC50 model, the R^2 is 0.66 (Figure 3.17) which is greater than that of EE% model and DOX NP IC50 model, showing better goodness-of-fit of the linear regression model. In Table 3.8, the highest bias between predicted and observed IC50 is 40.65% at testing point DTX 40_75_20k NP. The discrepancy is almost half of the observed IC50. The normalized-RMSE of 0.296 (Table 3.6) in the DOX NP IC50 testing set is higher than that of 0.206 (Table 3.9) in DTX NP IC50 model, and is the highest among all three models. Therefore, DOX NP IC50 model has the lowest prediction accuracy. The length scale of 2 in DTX NP IC50 model (Table 3.10) is higher than that of 1.54 in DOX NP IC50 model (Table 3.7), indicating that the kernel smoothness in DTX NP IC50 model is higher. Specifically, larger length scale means that the correlation between two IC50 values decreases at a greater distance between input data. It is worth mentioning that the length scales between EE% model and IC50 models are not comparable due to the different input space dimensionalities, where EE% model has 4 input variables and IC50 models have 3 input variables.

3.3.4.2 IC50 3D surface response curve

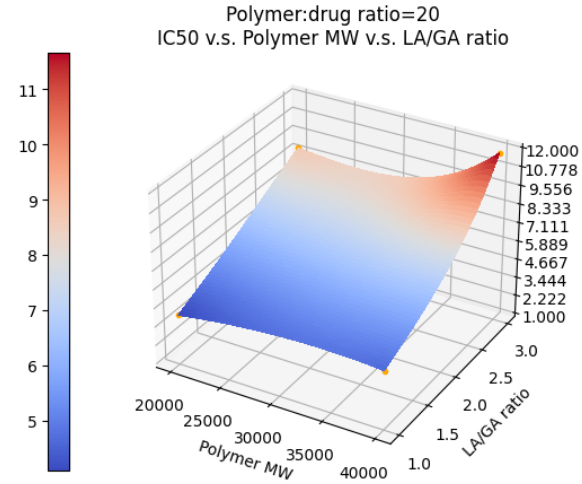
A.



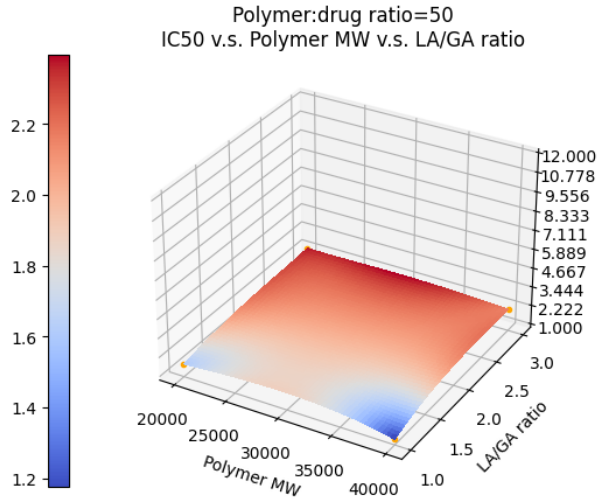
B.



C.



D.



E.

F.

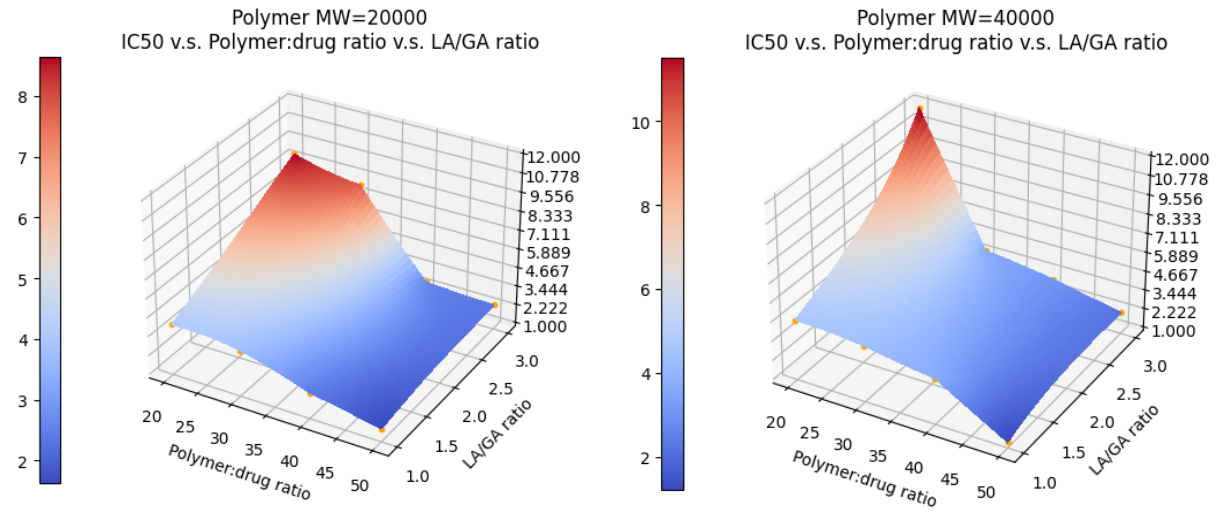
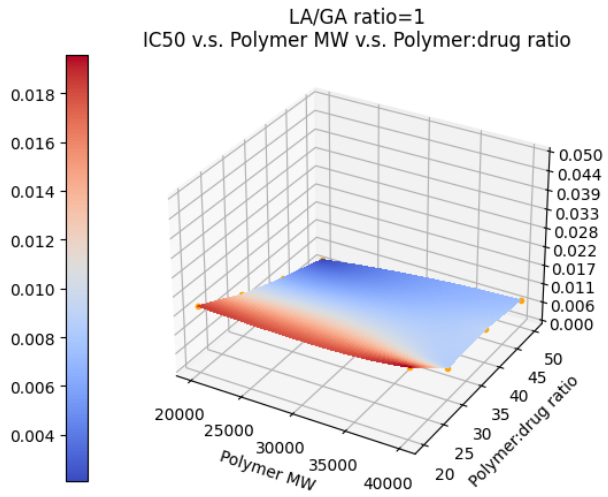


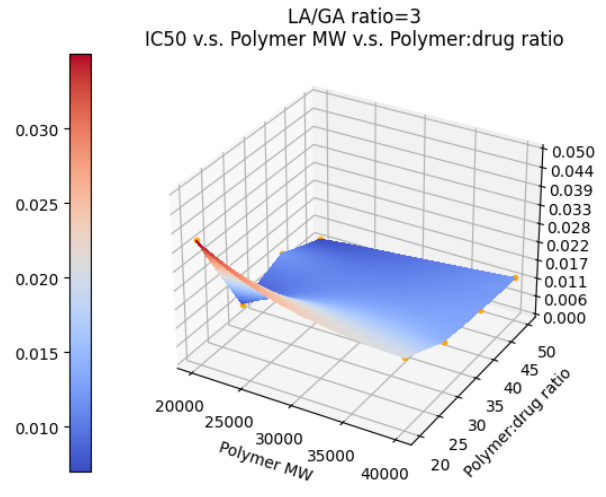
Figure 3.18. 3D surface response curve for DOX NP IC50 at two fixed levels of LA/GA ratio (1, 3), polymer: drug ratio (20, 50), and polymer MW (20000, 40000).

DOX NP IC50 curves were created at a fixed drug logP of 0.36. DOX NP IC50 barely varied with PLGA MW and LA/GA ratio when PLGA: drug ratio is at the upper level 50:1 in Figure 3.18D. However, at 20:1 PLGA: drug ratio, DOX NP IC50 increases by approximately 6μg/mL as LA/GA ratio increases in Figure 3.18C. More importantly, DOX NP IC50 follows an obvious increasing trend as PLGA: drug ratio decreases in Figure 3.18A, B, E, and F, opposite from DOX EE% variations in Figure 3.15M, N, U, and V. The reasons were explained in Section 3.2.1.1 and 3.2.2.1 regarding drug EE%, release, and accumulation in cells.

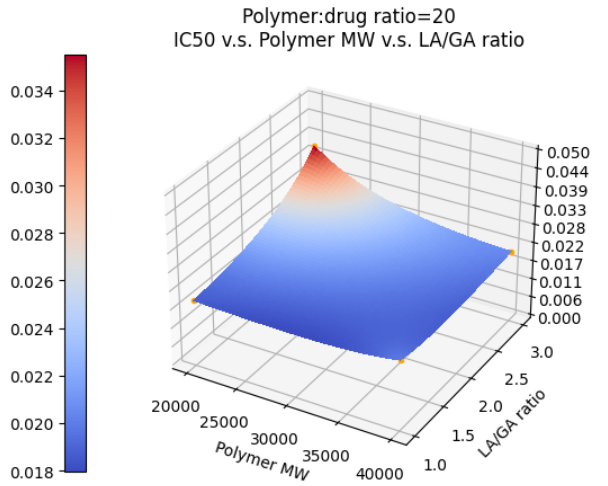
A. B.


LA/GA ratio=1
IC50 v.s. Polymer MW v.s. Polymer:drug ratio
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.050
0.044
0.039
0.033
0.028
0.022
0.017
0.011
0.006
0.000
20000
25000
30000
35000
40000
Polymer MW
20
25
30
35
40
45
50
Polymer:drug ratio

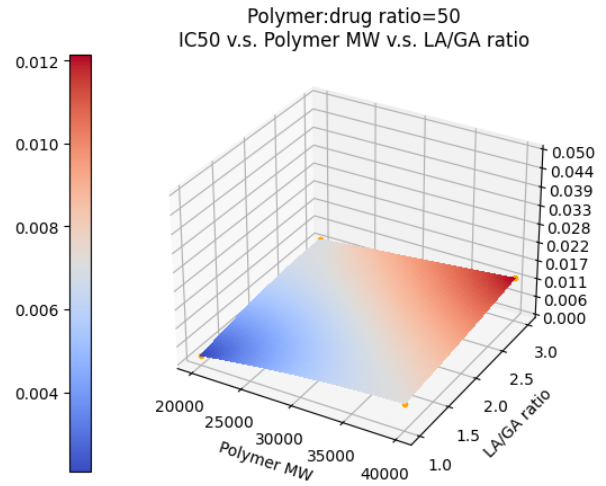
C.


LA/GA ratio=3
IC50 v.s. Polymer MW v.s. Polymer:drug ratio
0.030
0.025
0.020
0.015
0.010
0.050
0.044
0.039
0.033
0.028
0.022
0.017
0.011
0.006
0.000
20000
25000
30000
35000
40000
Polymer MW
20
25
30
35
40
45
50
Polymer:drug ratio

D.


Polymer:drug ratio=20
IC50 v.s. Polymer MW v.s. LA/GA ratio
0.034
0.032
0.030
0.028
0.026
0.024
0.022
0.020
0.018
0.050
0.044
0.039
0.033
0.028
0.022
0.017
0.011
0.006
0.000
20000
25000
30000
35000
40000
Polymer MW
1.0
1.5
2.0
2.5
3.0
LA/GA ratio

E.


Polymer:drug ratio=50
IC50 v.s. Polymer MW v.s. LA/GA ratio
0.012
0.010
0.008
0.006
0.004
0.050
0.044
0.039
0.033
0.028
0.022
0.017
0.011
0.006
0.000
20000
25000
30000
35000
40000
Polymer MW
1.0
1.5
2.0
2.5
3.0
LA/GA ratio

F.

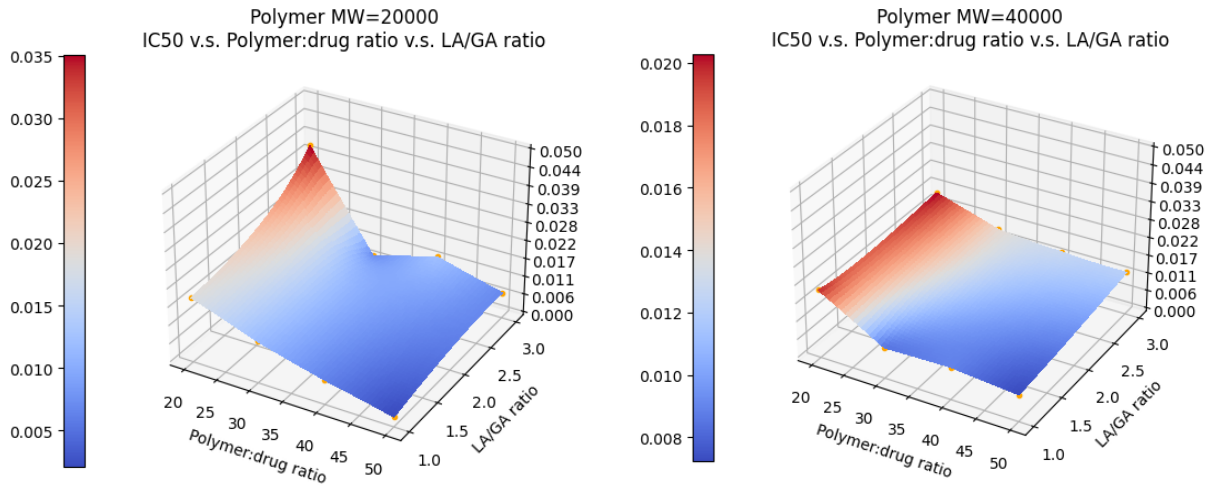


Figure 3.19. 3D surface response curve for DTX NP IC50 at two fixed levels of LA/GA ratio (1, 3), polymer: drug ratio (20, 50), and polymer MW (20000, 40000).

DTX NP IC50 curves were created at a fixed drug logP of 2.74. Similar to the surface trends of DOX NP IC50 in Figure 3.18, DTX IC50 also has a generally negative relationship with PLGA: drug ratio in Figure 3.19A, B, E, and F, except for a concave surface shape when PLGA: drug ratio is at the midpoint. Similar to DOX NPs, DTX NP IC50 does not change much with PLGA MW and PLGA LA/GA ratio when PLGA: drug ratio is at 50:1 (Figure 3.19D). Different from DOX NP IC50 which is higher at higher PLGA MW and LA/GA ratio in Figure 3.18C, DTX NP IC50 in Figure 3.19C is higher at lower PLGA MW as LA/GA ratio increases, when PLGA: drug ratio is fixed at 20:1. The difference might be due to the different drug logP that led to different drug-polymer interaction, resulting in different drug release profile and cellular accumulation.

3.4 BO model

The package 'bayes_opt' from Fernando N. was used to create BO model[114]. The objective function was set to 'none' as we are using the surrogate GP model to estimate it. Since no GP updates will be done in this project, BO model only serves as a non-iterated model to suggest x variable combination that maximizes the function value in EE% GP model and minimize that in DOX and DTX NP IC50 GP model. All data (both training and testing data) were used in the surrogate GP model in BO to make the best use of all available information of the objective function. The search space in each model was set between the lower and upper values for each x variable, that are 0.36 and 2.74 for drug logP, 20000 and 40000 for PLGA MW, 20 and 50 for PLGA: drug ratio, and 1 and 3 for LA/GA ratio. The 'UtilityFunction' serves as the acquisition function to recommend suggested x variable values on its local maxima with a hyperparameter 'kappa' that adjusts the trade-off between exploitation and exploration. In cases that updating GP model is required, 'kappa' can be set either high to explore areas where contain high uncertainty, or low to exploit known areas with less uncertainty but might also results in lower rewards. In exploration, the unexplored region is estimated and the corresponding uncertainty gets reduced. In exploitation, regions with higher reward (maximize/minimize objective function) will be exploited more. The suggested four physicochemical property values should be tested through *in-vitro* experiments for further GP model update and the next iteration in BO. Yet, due to time limitation, additional experiments have not been conducted.

In this study, 'kappa' was set to 0 (i.e., exploitation only) for the reason that GP model will not be updated with new data. The BO output values of x variables are summarized in Table 3.11. These formulations give the maximum drug EE% or minimum drug IC50 according to the known knowledge of GP-estimated true functions. In market, there are existing PLGA polymers at 55:45,

60:40, and 65:35 LA/GA ratio that are approximately the same as the recommended PLGA LA/GA ratios.

Table 3.11. Summary of BO-suggested x variable values to potentially optimize EE% and IC50.

Models	PLGA MW	PLGA LA/GA ratio	PLGA: drug ratio	Drug logP
EE%	20001.45	1.17 (55:45)	47.08	2.12
DOX NP IC50	39997.25	1.53 (60:40)	34.05	N/A
DTX NP IC50	20007.87	1.70 (65:35)	45.19	N/A

Chapter 4. Conclusions and future directions

4.1 Conclusions

The effects of NP compositions on NP biological behaviors are considerably important, as they are associated with drug-NP composition interactions, cell entry, cellular uptake, and drug release. The effort of selecting the correct material properties and conducting *in-vitro* response experiments can be huge. With the aim of efficiently optimizing NP materials in achieving higher drug loading and therapeutic effects, we utilized ML-based prediction models to map the function of inputs (NP material physicochemical properties) and outputs (EE% and IC50). ML techniques have successfully accelerated drug design process, reported by many papers in the past decade.

As a proof-of-concept, we formulated 32 drug-loaded NPs and used GP and BO models to predict and optimize the NP EE% and IC50 under different PLGA MW, PLGA LA/GA ratio, PLGA: drug ratio, and drug logP in OVCAR3 treatment. We analyzed the possible relationships between the four NP physicochemical properties and EE% or IC50, respectively, via statistical analysis. Analysis was done separately for each drug, considering their different logP. Thereinto, DOX EE% generally increases as LA/GA ratio increases when PLGA MW is at 20kDa, except for formulations that are at the highest PLGA: drug ratio. However, the trend reverses when PLGA MW is at 40kDa. DTX EE% mostly increases as LA/GA ratio increases when PLGA MW is at 20kDa, except for NPs with 30:1 PLGA: drug ratio. But no relationship was found when PLGA MW is at 40kDa. In both 50:50 and 75:25 PLGA NPs, DOX EE% is inversely proportional to PLGA MW. No relationship was observed for DTX EE%. For both drugs, their EE% increase with PLGA: drug ratio, except for DTX NPs with 50:50 and 75:25 20kDa PLGA at 50:1 PLGA: drug ratio that might have experienced PLGA saturation. Other than DTX NP with 50:50 20kDa PLGA at 50:1 PLGA: drug ratio, the rest all have higher EE% than DOX NPs. No obvious trend can be

concluded between DOX NP IC50 and LA/GA ratio or PLGA MW. However, 75:25 40kDa PLGA NPs at 20:1 and 50:1 PLGA: drug ratio have higher IC50 than 50:50 40kDa PLGA NPs. DTX NP IC50 increases with LA/GA ratio in 20kDa PLGA NPs, except for NPs at 30:1 PLGA: drug ratio. The same trend was observed in 40kDa PLGA NPs that are at 30:1 and 50:1 PLGA: drug ratio. In DTX-loaded 50:50 PLGA NPs, only 40kDa PLGA NP at 50:1 PLGA: drug ratio has higher IC50 than 20kDa PLGA NP. No clear relationship between IC50 and PLGA MW was found in DTX-loaded 75:25 PLGA NPs. Both DOX and DTX-loaded NPs have a decreasing IC50 as PLGA: drug ratio increases. In summary, it is difficult to conclude uniform relationships between material properties and responses, other than the positive relationship between NP EE% and PLGA: drug ratio, and negative relationship between NP IC50 and PLGA: drug ratio for both drugs.

We have also successfully established one EE% and two IC50 GP models upon data training and testing, with normalized-RMSE (testing set) of 0.187 in EE% model, 0.296 in DOX IC50 model, and 0.206 in DTX IC50 model. 3D surface response curves were generated to visualize the model prediction, with two x variables on the horizontal plane and EE% or IC50 on the vertical plane. However, due to insufficient data, the prediction surfaces show limited information. More data is necessary to display more informatic and accurate prediction surface response curves. Due to time limitation, additional experiments were not performed to update the GP surrogate model and fully BO was not fully utilized in balancing exploration/exploitation to suggest the next valuable formulation for evaluation. A simplified BO model was built to output the suggested predictor variable values that gives the maximum EE% or minimum IC50 in the current GP-estimated model, considering exploitation only. Under such circumstance, the output formulation can be considered as the current recommended formulation as well as the next valuable formulation that needs to be tested on since areas with uncertainty is not explored. In other words,

these formulations are the optimal choices, given the current training data and true function information we have, but are not necessarily the best formulations when true function is explored more and more in the future. The recommended NP formulation that can achieve the highest EE% is loaded with drug with logP of 2.12 and 47_55_20001 PLGA. DOX NP with 34_60_39997 PLGA and DTX NP with 45_65_20008 PLGA were recommended to achieve the lowest IC50 in each of the model. If we considered both exploration and exploitation in the BO model, the BO-suggested formulation might not result in the highest or lowest function value but would be the next one that is worth testing to potentially help to find the ultimate optimal formulation until BO iteration stops at certain stopping criteria.

To improve the model prediction accuracy, many approaches can be considered. That includes increasing training data set with larger diversity, choosing a more appropriate kernel function, and tuning the hyperparameters in the kernel function. The choice of kernel depends on the behaviors and structures of the dataset itself, therefore, different datasets need different kernels for better fitting.

This project can be extended to other applications differently as a template. If one is also interested in OVCAR3 cells and input and output variables that were specifically used in this project, this GP model platform could predict the corresponding NP EE% and IC50 of the new input data. Restricted by the definition of GP where prediction relies on the similarities between data points, failure of generalization may occur if the new data domain is far from the training data domain a lot. The solution to this could be expanding the training data domain by introducing a variety of data in. If the cell line, input variable features, output variable features, or all are changed, the GP model structure still applies but the kernel and its hyperparameter shall be updated to be

suitable for specific tasks. Also, the variable features of the testing data should be the same as that of training data to ensure a successful model training and application.

In conclusion, we proved the hypothesis that the GP model can make relatively accurate predictions, but we failed to prove the hypothesis that BO optimizes formulations due to limited time of performing additional experiments. GP has shown its potential in accelerating NP formulation development by making predictions on response variables without conducting wet-lab experiments. This ML technique has shown its power in making rapid predictions in order to reduce costs and time. Further improvements and implementations are worth to be done to fully leverage ML in boosting general NP formulation development.

4.2 Limitations

Due to time limitations, each experiment was only repeated once with at least triplicate measurements. In preliminary experiments, two NPs were made for 3 different times with triplicate measurements in characterization and EE% testing, shown in Appendix A Table A2 to demonstrate that NPs were reproducible. Furthermore, the uncertainty of response predictions was not included due to time limitation. Uncertainty estimation as GP's nature may provide more interpretation on model predictability and more accurate relationships between predicted and observed responses in linear regression graphs. Additionally, measurement uncertainty has not been included in data training. There is necessity to add this data noise to generate more comprehensive prediction.

Due to the limited number of data, the 3D prediction surface response curves lack details in between the lower and upper levels of each x variable. Specifically, since PLGA MW, PLGA LA/GA ratio, and logP variables only have two input values in each, and PLGA: drug ratio only has four input values, it is challenging to observe any other trends of EE% or IC50 at other points

within these domains. The biases between predicted y and observed y are relatively high in IC50 models, with one of the DOX IC50 bias reached 91% and one of the DTX IC50 bias at 40.65%. In contrast, the highest bias in EE% model is only 28.73% which is much lower. This is due to the limited number of data in each IC50 training set that is only half of the size of EE% training set.

4.3 Future directions

NP drug release and cellular uptake studies can provide comprehensive examination on how the NP components play roles in effecting NP EE% and IC50. Therefore, it would be helpful to include these studies for us to explicitly explain the NP behavior.

In the current model, only four predictors were analyzed for drug EE% and three were analyzed for drug IC50. However, there are many other predictors that could potentially influence the two responses, such as NP size and zeta potential. In future work, with the addition of more predictors, the model estimation and responses prediction might be improved with richer information from different aspects.

To enhance the model prediction accuracy, more data should be generated and possibly include more predictor variables for deeper evaluation on the NP compositions. Another method to improve model prediction accuracy is leave-one-out cross validation (LOO CV). N-1 data points should be used to train the model, leaving one data to test the model performance. This process iterates N times (number of data) until each of the data has been ran as test set. This allows a more comprehensive data training, making the most use of the limited data. With increased number of training data in comparison to conventional training/testing data split method, the model prediction accuracy can be improved markedly.

In the case of a lack of data, a statistical strategy of bootstrapping could be used to simulate data from the present small dataset[115]. This resampling process allows estimation of the sampling distribution. Each time, n samples are randomly resampled from the original dataset (n samples), and form a bootstrap dataset. Data can be resampled multiple times. Bootstrap can provide various statistics and metrics of the estimates of quantity of interest[115]. This method enables better understanding of the data behavior and distribution in population.

References

1. Adepu, S. & Ramakrishna, S. Controlled Drug Delivery Systems: Current Status and Future Directions. *Molecules* **26**, 5905 (2021).
2. Maheshwari, R. *et al.* Chapter 20 - Manipulation of Physiological Processes for Pharmaceutical Product Development. in *Dosage Form Design Considerations* (ed. Tekade, R. K.) 701–729 (Academic Press, 2018). doi:10.1016/B978-0-12-814423-7.00020-4.
3. Scheller, E. L. & Krebsbach, P. H. Gene Therapy. *J Dent Res* **88**, 585–596 (2009).
4. Chen, J., Guo, Z., Tian, H. & Chen, X. Production and clinical development of nanoparticles for gene delivery. *Molecular Therapy - Methods & Clinical Development* **3**, (2016).
5. Hassan, S. S. M., Kamel, A. H., Hashem, H. M. & Bary, E. M. A. Drug delivery systems between metal, liposome, and polymer-based nanomedicine: A review. *ECB* **9**, 91 (2020).
6. Rizvi, S. A. A. & Saleh, A. M. Applications of nanoparticle systems in drug delivery technology. *Saudi Pharm J* **26**, 64–70 (2018).
7. De Jong, W. H. & Borm, P. J. Drug delivery and nanoparticles: Applications and hazards. *Int J Nanomedicine* **3**, 133–149 (2008).
8. Pardridge, W. M. Drug transport across the blood–brain barrier. *J Cereb Blood Flow Metab* **32**, 1959–1972 (2012).
9. Mitchell, M. J. *et al.* Engineering precision nanoparticles for drug delivery. *Nat Rev Drug Discov* **20**, 101–124 (2021).
10. Vandervoort, J. & Ludwig, A. Biocompatible stabilizers in the preparation of PLGA nanoparticles: a factorial design study. *International Journal of Pharmaceutics* **238**, 77–92 (2002).

11. Ibrahim, A. M., Alzahrani, H. A. A., Abd El-Latif, M. M. & Selim, M. M. Influence of different stabilizers on the morphology of gold nanoparticles. *Bulletin of the National Research Centre* **43**, 33 (2019).

12. Sivaram, A. J., Wardiana, A., Howard, C. B., Mahler, S. M. & Thurecht, K. J. Recent Advances in the Generation of Antibody–Nanomaterial Conjugates. *Advanced Healthcare Materials* **7**, 1700607 (2018).

13. Gavas, S., Quazi, S. & Karpiński, T. M. Nanoparticles for Cancer Therapy: Current Progress and Challenges. *Nanoscale Res Lett* **16**, 173 (2021).

14. Roy, P. S. & Saikia, B. J. Cancer and cure: A critical analysis. *Indian Journal of Cancer* **53**, 441 (2016).

15. Rasool, M. *et al.* New challenges in the use of nanomedicine in cancer therapy. *Bioengineered* **13**, 759–773 (2022).

16. Liu, Y. *et al.* Nanoparticles advanced from preclinical studies to clinical trials for lung cancer therapy. *Cancer Nanotechnol* **14**, 28 (2023).

17. Anselmo, A. C. & Mitragotri, S. Nanoparticles in the clinic. *Bioengineering & Translational Medicine* **1**, 10–29 (2016).

18. Miao, L. & Huang, L. Exploring the Tumor Microenvironment with Nanoparticles. *Cancer Treat Res* **166**, 193–226 (2015).

19. Nakamura, Y., Mochida, A., Choyke, P. L. & Kobayashi, H. Nano-drug delivery: Is the enhanced permeability and retention (EPR) effect sufficient for curing cancer? *Bioconjug Chem* **27**, 2225–2238 (2016).

20. Donahue, N. D., Acar, H. & Wilhelm, S. Concepts of nanoparticle cellular uptake, intracellular trafficking, and kinetics in nanomedicine. *Advanced Drug Delivery Reviews* **143**, 68–96 (2019).

21. Rennick, J. J., Johnston, A. P. R. & Parton, R. G. Key principles and methods for studying the endocytosis of biological and nanoparticle therapeutics. *Nat. Nanotechnol.* **16**, 266–276 (2021).

22. Means, N., Elechalawar, C. K., Chen, W. R., Bhattacharya, R. & Mukherjee, P. Revealing macropinocytosis using nanoparticles. *Molecular Aspects of Medicine* **83**, 100993 (2022).

23. Manzanares, D. & Ceña, V. Endocytosis: The Nanoparticle and Submicron Nanocompounds Gateway into the Cell. *Pharmaceutics* **12**, 371 (2020).

24. Smith, S. A., Selby, L. I., Johnston, A. P. R. & Such, G. K. The Endosomal Escape of Nanoparticles: Toward More Efficient Cellular Delivery. *Bioconjugate Chem.* **30**, 263–272 (2019).

25. Ahmad, A., Khan, J. M. & Haque, S. Strategies in the design of endosomolytic agents for facilitating endosomal escape in nanoparticles. *Biochimie* **160**, 61–75 (2019).

26. Smith, M. C., Crist, R. M., Clogston, J. D. & McNeil, S. E. Zeta potential: a case study of cationic, anionic, and neutral liposomes. *Anal Bioanal Chem* **409**, 5779–5787 (2017).

27. Freeman, E. C., Weiland, L. M. & Meng, W. S. Modeling the Proton Sponge Hypothesis: Examining Proton Sponge Effectiveness for Enhancing Intracellular Gene Delivery through Multiscale Modeling. *J Biomater Sci Polym Ed* **24**, 398–416 (2013).

28. Tracey, S. R., Smyth, P., Barelle, C. J. & Scott, C. J. Development of next generation nanomedicine-based approaches for the treatment of cancer: we've barely scratched the surface. *Biochemical Society Transactions* **49**, 2253–2269 (2021).

29. Ma, D. Enhancing endosomal escape for nanoparticle mediated siRNA delivery. *Nanoscale* **6**, 6415 (2014).

30. Fröhlich, E. The role of surface charge in cellular uptake and cytotoxicity of medical nanoparticles. *Int J Nanomedicine* **7**, 5577–5591 (2012).

31. Namiot, E. D., Sokolov, A. V., Chubarev, V. N., Tarasov, V. V. & Schiöth, H. B. Nanoparticles in Clinical Trials: Analysis of Clinical Trials, FDA Approvals and Use for COVID-19 Vaccines. *Int J Mol Sci* **24**, 787 (2023).

32. Zielińska, A. *et al.* Polymeric Nanoparticles: Production, Characterization, Toxicology and Ecotoxicology. *Molecules* **25**, 3731 (2020).

33. Patra, J. K. *et al.* Nano based drug delivery systems: recent developments and future prospects. *Journal of Nanobiotechnology* **16**, 71 (2018).

34. Lu, H., Zhang, S., Wang, J. & Chen, Q. A Review on Polymer and Lipid-Based Nanocarriers and Its Application to Nano-Pharmaceutical and Food-Based Systems. *Front. Nutr.* **8**, 783831 (2021).

35. Patel, T., Zhou, J., Piepmeier, J. M. & Saltzman, W. M. Polymeric Nanoparticles for Drug Delivery to the Central Nervous System. *Adv Drug Deliv Rev* **64**, 701–705 (2012).

36. Alsaheb, R. A. A. *et al.* Recent applications of polylactic acid in pharmaceutical and medical industries. (2015).

37. Gatoo, M. A. *et al.* Physicochemical Properties of Nanomaterials: Implication in Associated Toxic Manifestations. *Biomed Res Int* **2014**, 498420 (2014).

38. Makadia, H. K. & Siegel, S. J. Poly Lactic-co-Glycolic Acid (PLGA) as Biodegradable Controlled Drug Delivery Carrier. *Polymers (Basel)* **3**, 1377–1397 (2011).

39. You, X. *et al.* Effects of polymer molecular weight on in vitro and in vivo performance of nanoparticle drug carriers for lymphoma therapy. *Chinese Chemical Letters* **34**, 107720 (2023).

40. Song, X. *et al.* Dual agents loaded PLGA nanoparticles: Systematic study of particle size and drug entrapment efficiency. *European Journal of Pharmaceutics and Biopharmaceutics* **69**, 445–453 (2008).

41. Song, X. *et al.* PLGA nanoparticles simultaneously loaded with vincristine sulfate and verapamil hydrochloride: Systematic study of particle size and drug entrapment efficiency. *International Journal of Pharmaceutics* **350**, 320–329 (2008).

42. Amasya, G., Badilli, U., Aksu, B. & Tarimci, N. Quality by design case study 1: Design of 5-fluorouracil loaded lipid nanoparticles by the W/O/W double emulsion — Solvent evaporation method. *European Journal of Pharmaceutical Sciences* **84**, 92–102 (2016).

43. Ram, A. & Kadim, A. Shear degradation of polymer solutions. *Journal of Applied Polymer Science* **14**, 2145–2156 (1970).

44. Barichello, J. M., Morishita, M., Takayama, K. & Nagai, T. Encapsulation of Hydrophilic and Lipophilic Drugs in PLGA Nanoparticles by the Nanoprecipitation Method. *Drug Development and Industrial Pharmacy* **25**, 471–476 (1999).

45. Li, Q., Li, X. & Zhao, C. Strategies to Obtain Encapsulation and Controlled Release of Small Hydrophilic Molecules. *Frontiers in Bioengineering and Biotechnology* **8**, (2020).

46. Ramazani, F. *et al.* Strategies for encapsulation of small hydrophilic and amphiphilic drugs in PLGA microspheres: State-of-the-art and challenges. *International Journal of Pharmaceutics* **499**, 358–367 (2016).

47. Zhao, G. *et al.* Effect of Carrier Lipophilicity and Preparation Method on the Properties of Andrographolide–Solid Dispersion. *Pharmaceutics* **11**, 74 (2019).

48. Zhuang, B., Ramanauskaite, G., Koa, Z. Y. & Wang, Z.-G. Like dissolves like: A first-principles theory for predicting liquid miscibility and mixture dielectric constant. *Sci Adv* **7**, eabe7275 (2021).

49. Anselmo, A. C. & Mitragotri, S. Nanoparticles in the clinic: An update. *Bioeng Transl Med* **4**, (2019).

50. Murthy, S. K. Nanoparticles in modern medicine: State of the art and future challenges. *Int J Nanomedicine* **2**, 129–141 (2007).

51. Levasseur, L. M., Slocum, H. K., Rustum, Y. M. & Greco, W. R. Modeling of the time-dependency of in vitro drug cytotoxicity and resistance. *Cancer Res* **58**, 5749–5761 (1998).

52. Gardner, S. N. A Mechanistic, Predictive Model of Dose-Response Curves for Cell Cycle Phase-specific and -nonspecific Drugs1. *Cancer Research* **60**, 1417–1425 (2000).

53. Berrouet, C., Dorilas, N., Rejniak, K. A. & Tuncer, N. *Comparison of drug inhibitory effects (IC $_{50}$) in monolayer and spheroid cultures*. http://biorxiv.org/lookup/doi/10.1101/2020.05.05.079285 (2020) doi:10.1101/2020.05.05.079285.

54. Reid, B. M., Permuth, J. B. & Sellers, T. A. Epidemiology of ovarian cancer: a review. *Cancer Biol Med* **14**, 9–32 (2017).

55. Lheureux, S., Gourley, C., Vergote, I. & Oza, A. M. Epithelial ovarian cancer. *The Lancet* **393**, 1240–1253 (2019).

56. Matulonis, U. A. *et al.* Ovarian cancer. *Nat Rev Dis Primers* **2**, 16061 (2016).

57. Lisio, M.-A., Fu, L., Goyeneche, A., Gao, Z. & Telleria, C. High-Grade Serous Ovarian Cancer: Basic Sciences, Clinical and Therapeutic Standpoints. *Int J Mol Sci* **20**, 952 (2019).

58. Yao, Y. *et al.* Nanoparticle-Based Drug Delivery in Cancer Therapy and Its Role in Overcoming Drug Resistance. *Frontiers in Molecular Biosciences* **7**, (2020).

59. Rafiyath, S. M. *et al.* Comparison of safety and toxicity of liposomal doxorubicin vs. conventional anthracyclines: a meta-analysis. *Exp Hematol Oncol* **1**, 10 (2012).

60. Singal, P. K., Li, T., Kumar, D., Danelisen, I. & Iliskovic, N. Adriamycin-induced heart failure: mechanism and modulation.

61. Barenholz, Y. (Chezy). Doxil® — The first FDA-approved nano-drug: Lessons learned. *Journal of Controlled Release* **160**, 117–134 (2012).

62. Thorn, C. F. *et al.* Doxorubicin pathways: pharmacodynamics and adverse effects. *Pharmacogenetics and Genomics* **21**, 440–446 (2011).

63. Kuroki, L. & Guntupalli, S. R. Treatment of epithelial ovarian cancer. *BMJ* m3773 (2020) doi:10.1136/bmj.m3773.

64. Gordon, A. N. *et al.* Phase II study of liposomal doxorubicin in platinum- and paclitaxel-refractory epithelial ovarian cancer. *J Clin Oncol* **18**, 3093–3100 (2000).

65. Katsumata, N. Docetaxel: an alternative taxane in ovarian cancer. *Br J Cancer* **89**, S9–S15 (2003).

66. Kaur, R., Kaur, G., Gill, R. K., Soni, R. & Bariwal, J. Recent developments in tubulin polymerization inhibitors: An overview. *European Journal of Medicinal Chemistry* **87**, 89–124 (2014).

67. Sumera, Anwar, A., Ovais, M., Khan, A. & Raza, A. Docetaxel-loaded solid lipid nanoparticles: a novel drug delivery system. *IET Nanobiotechnol* **11**, 621–629 (2017).

68. Vasey, P. A. *et al.* Docetaxel-carboplatin as first line chemotherapy for epithelial ovarian cancer. *Br J Cancer* **84**, 170–178 (2001).

69. Farha, N. G. & Kasi, A. Docetaxel. in *StatPearls* (StatPearls Publishing, 2023).

70. Ghassami, E. *et al.* Pharmacokinetics and in vitro/in vivo antitumor efficacy of aptamer-targeted Ecoflex® nanoparticles for docetaxel delivery in ovarian cancer. *Int J Nanomedicine* **13**, 493–504 (2018).

71. Janiesch, C., Zschech, P. & Heinrich, K. Machine learning and deep learning. *Electron Markets* **31**, 685–695 (2021).

72. Kolluri, S., Lin, J., Liu, R., Zhang, Y. & Zhang, W. Machine Learning and Artificial Intelligence in Pharmaceutical Research and Development: a Review. *AAPS J* **24**, 19 (2022).

73. Jariwala, N. *et al.* Intriguing of pharmaceutical product development processes with the help of artificial intelligence and deep/machine learning or artificial neural network. *Journal of Drug Delivery Science and Technology* **87**, 104751 (2023).

74. Gupta, R. *et al.* Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Mol Divers* **25**, 1315–1360 (2021).

75. Ekins, S. The Next Era: Deep Learning in Pharmaceutical Research. *Pharm Res* **33**, 2594–2603 (2016).

76. Abdulla, A. *et al.* Project IDentif.AI: Harnessing Artificial Intelligence to Rapidly Optimize Combination Therapy Development for Infectious Disease Intervention. *Adv. Therap.* **3**, 2000034 (2020).

77. Crampon, K., Giorkallos, A., Deldossi, M., Baud, S. & Steffenel, L. A. Machine-learning methods for ligand–protein molecular docking. *Drug Discovery Today* **27**, 151–164 (2022).

78. Khamis, M. A., Gomaa, W. & Ahmed, W. F. Machine learning in computational docking. *Artificial Intelligence in Medicine* **63**, 135–152 (2015).

79. Sapoval, N. *et al.* Current progress and open challenges for applying deep learning across the biosciences. *Nat Commun* **13**, 1728 (2022).

80. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nat Rev Genet* **16**, 321–332 (2015).

81. Turzo, S. B. A., Hantz, E. R. & Lindert, S. Applications of machine learning in computer-aided drug discovery. *QRB Discov* **3**, e14 (2022).

82. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* **18**, 1196–1203 (2021).

83. Wang, Y., Carter, B. Z., Li, Z. & Huang, X. Application of machine learning methods in clinical trials for precision medicine. *JAMIA Open* **5**, ooab107 (2022).

84. Sarker, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN COMPUT. SCI.* **2**, 160 (2021).

85. Bannigan, P. *et al.* Machine learning models to accelerate the design of polymeric long-acting injectables. *Nat Commun* **14**, 35 (2023).

86. Bannigan, P. *et al.* Machine learning directed drug formulation development. *Advanced Drug Delivery Reviews* **175**, 113806 (2021).

87. Ameli, S. & Shadden, S. C. Noise Estimation in Gaussian Process Regression. Preprint at http://arxiv.org/abs/2206.09976 (2022).

88. Mchutchon, A. & Rasmussen, C. E. Gaussian Process Training with Input Noise.

89. Kanagawa, M., Hennig, P., Sejdinovic, D. & Sriperumbudur, B. K. Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences. Preprint at https://doi.org/10.48550/arXiv.1807.02582 (2018).

90. Hensman, J., Fusi, N. & Lawrence, N. D. Gaussian Processes for Big Data.

91. Genton, M. G. Classes of Kernels for Machine Learning: A Statistics Perspective. *Journal of Machine Learning Research* **2**, 299–312 (2001).

92. Noack, M. M. & Sethian, J. A. Advanced Stationary and Non-Stationary Kernel Designs for Domain-Aware Gaussian Processes. *Commun. Appl. Math. Comput. Sci.* **17**, 131–156 (2022).

93. Rasmussen, C. E. & Williams, C. K. I. *Gaussian processes for machine learning*. (MIT Press, 2006).

94. Bahg, G., Evans, D. G., Galdo, M. & Turner, B. M. Gaussian process linking functions for mind, brain, and behavior. *Proceedings of the National Academy of Sciences* **117**, 29398–29406 (2020).

95. Beckers, T. An Introduction to Gaussian Process Models. Preprint at https://doi.org/10.48550/arXiv.2102.05497 (2021).

96. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

97. Strober, W. Trypan Blue Exclusion Test of Cell Viability. *Curr Protoc Immunol* **111**, A3.B.1-A3.B.3 (2015).

98. Bilati, U., Allémann, E. & Doelker, E. Sonication Parameters for the Preparation of Biodegradable Nanocapsulesof Controlled Size by the Double Emulsion Method. *Pharmaceutical Development and Technology* **8**, 1–9 (2003).

99. Wang, Z., Lin, X., Rui, Z., Xu, M. & Zhan, S. The Role of Shearing Energy and Interfacial Gibbs Free Energy in the Emulsification Mechanism of Waxy Crude Oil. *Energies* **10**, 721 (2017).

100. Tadros, T. F. Emulsion Formation, Stability, and Rheology. in *Emulsion Formation and Stability* 1–75 (John Wiley & Sons, Ltd, 2013). doi:10.1002/9783527647941.ch1.

101. Murakami, H. *et al.* Influence of the degrees of hydrolyzation and polymerization of poly(vinylalcohol) on the preparation and properties of poly(dl-lactide-co-glycolide) nanoparticle. *International Journal of Pharmaceutics* **149**, 43–49 (1997).

102. Daina, A., Michielin, O. & Zoete, V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci Rep* **7**, 42717 (2017).

103. Stipa, P. *et al.* Prediction of drug-carrier interactions of PLA and PLGA drug-loaded nanoparticles by molecular dynamics simulations. *European Polymer Journal* **147**, 110292 (2021).

104. Hu, H. *et al.* Entanglements in marginal solutions: a means of tuning pre-aggregation of conjugated polymers with positive implications for charge transport. *Journal of Materials Chemistry C* **3**, 7394–7404 (2015).

105. Presmanes, C. *et al.* Effect of PLGA hydrophilia on the drug release and the hypoglucemic activity of different insulin-loaded PLGA microspheres. *Journal of Microencapsulation* **28**, 791–798 (2011).

106. Ravi, S. *et al.* Development and characterization of polymeric microspheres for controlled release protein loaded drug delivery system. *Indian J Pharm Sci* **70**, 303 (2008).

107. Fonseca, C., Simões, S. & Gaspar, R. Paclitaxel-loaded PLGA nanoparticles: preparation, physicochemical characterization and in vitro anti-tumoral activity. *Journal of Controlled Release* **83**, 273–286 (2002).

108. GraphPad Prism 10 Curve Fitting Guide - Equation: Absolute IC50. https://www.graphpad.com/guides/prism/latest/curve-fitting/reg_absolute_ic50.htm.

109. Sadhukha, T. & Prabha, S. Encapsulation in Nanoparticles Improves Anti-cancer Efficacy of Carboplatin. *AAPS PharmSciTech* **15**, 1029–1038 (2014).

110. Shah, S. *et al.* Fluorescence properties of doxorubicin in PBS buffer and PVA films. *Journal of Photochemistry and Photobiology B: Biology* **170**, 65–69 (2017).

111. Blum, M. & Riedmiller, M. Optimization of Gaussian Process Hyperparameters using Rprop. *Computational Intelligence* (2013).

112. Verrelst, J., Rivera, J. P., Moreno, J. & Camps-Valls, G. Gaussian processes uncertainty estimates in experimental Sentinel-2 LAI and leaf chlorophyll content retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing* **86**, 157–167 (2013).

113. Liemohn, M. W. *et al.* RMSE is not enough: Guidelines to robust data-model comparisons for magnetospheric physics. *Journal of Atmospheric and Solar-Terrestrial Physics* **218**, 105624 (2021).

114. Nogueira, F. Bayesian Optimization: Open source constrained global optimization tool for Python. (2014).

115. Kulesa, A., Krzywinski, M., Blainey, P. & Altman, N. Sampling distributions and the bootstrap. *Nature Methods* **12**, 477–478 (2015).

Appendices

Appendix A

Table A1. Size and zeta potential of each NP formulation. Data represent mean ± SD (N=3).

NP	Size (nm)	Zeta potential (mV)
Blank 20_50_20k	220.4±1.6	-26.2±0.8
Blank 50_50_20k	236.1±2.9	-22.6±0.3
Blank 20_50_40k	235.6±4.1	-27.5±0.8
Blank 50_50_40k	249.4±2.8	-22.0±0.3
Blank 20_75_20k	225.0±4.4	-29.2±0.6
Blank 30_75_20k	244.8±3.7	-26.0±0.7
Blank 50_75_20k	256.8±2.1	-20.3±0.9
Blank 20_75_40k	228.4±5.5	-27.1±0.4
Blank 30_75_40k	252.0±5.4	-28.8±0.3
Blank 40_75_40k	259.5±4.3	-21.7±1.2
Blank 50_75_40k	263.5±1.6	-23.4±1.1
DOX 20_50_20k	214.4±3.8	-28.7±0.7
DOX 30_50_20k	230.9±1.7	-25.7±0.2
DOX 40_50_20k	235.3±3.6	-25.2±0.3
DOX 50_50_20k	249.0±2.5	-20.1±0.5
DOX 20_50_40k	221.8±1.4	-25.0±0.5
DOX 30_50_40k	241.3±2.3	-25.7±0.8

DOX 40_50_40k	247.5±2.9	-24.4±0.5
DOX 50_50_40k	239.6±3.1	-15.5±0.3
DOX 20_75_20k	223.1±2.5	-25.0±0.4
DOX 30_75_20k	227.5±5.2	-20.9±0.4
DOX 40_75_20k	249.0±3.9	-23.4±0.3
DOX 50_75_20k	261.7±5.1	-23.8±0.4
DOX 20_75_40k	232.7±2.7	-22.5±0.2
DOX 30_75_40k	234.9±1.1	-25.1±0.1
DOX 40_75_40k	258.8±2.0	-26.0±0.9
DOX 50_75_40k	255.4±3.5	-24.9±0.4
DTX 20_50_20k	208.9±3.4	-29.4±0.8
DTX 30_50_20k	235.5±2.2	-30.8±0.8
DTX 40_50_20k	244.8±5.0	-17.4±0.6
DTX 50_50_20k	242.1±4.4	-22.0±0.3
DTX 20_50_40k	233.6±3.8	-23.2±0.9
DTX 30_50_40k	244.3±5.2	-25.3±0.7
DTX 40_50_40k	257.8±6.0	-22.3±2.1
DTX 50_50_40k	257.9±6.7	-15.4±1.0
DTX 20_75_20k	235.9±4.8	-24.8±0.4
DTX 30_75_20k	250.6±2.1	-20.4±1.8
DTX 40_75_20k	255.4±0.8	-36.2±0.3
DTX 50_75_20k	264.6±4.7	-16.5±0.5

DTX 20_75_40k	230.1±4.9	-26.8±1.0
DTX 30_75_40k	243.6±1.1	-25.2±0.7
DTX 40_75_40k	251.4±3.5	-23.6±2.4
DTX 50_75_40k	267.7±1.5	-15.0±0.7

Table A2. Size, zeta potential, and drug EE% for DOX 20_75_40k and DTX 20_50_20k. DOX 20_75_40k was collected at 10,000 ×g. Each formulation was performed 3 different times with triplicate measurements. Data represent the mean ± SD (N=3).

	Size (nm)	**PDI**	**Zeta potential (mV)**	**EE%**
DOX 20_75_40k	243.77±3.07	0.023±0.015	-22.32±10.06	12.44±2.07
DTX 20_50_20k	215.93±2.45	0.023±0.002	-28.01±1.51	29.15±3.45

Appendix B

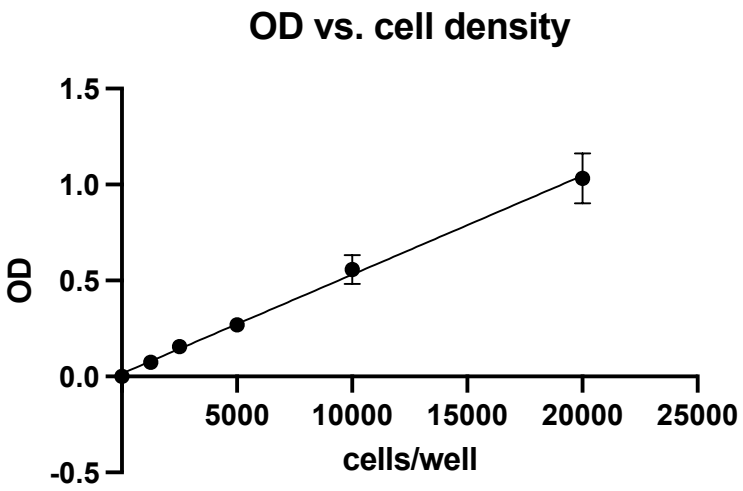


Figure B1. Absorbance versus cell density curve at wavelength 490 nm. Cells at different densities were seeded and equilibrated for 1.5 hrs before readings were taken. Data represent the mean ± SD (N=6). The result shows that the absorbance is positively proportional to the cell number with a high linear regression R^2 of 0.971.

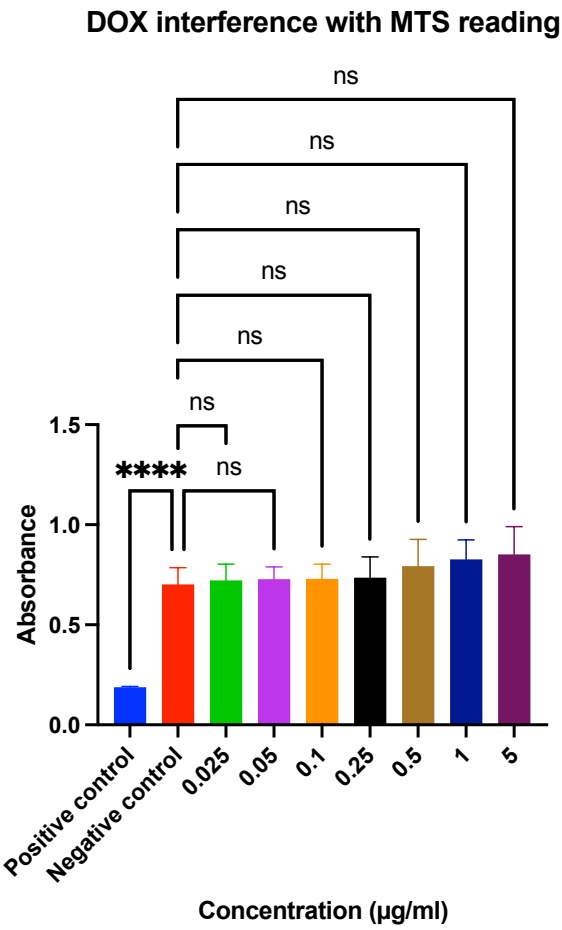


Figure B2. DOX interference with MTS assay in the presence of cells. 5000 cells were seeded into each well and incubated for 24 hrs prior to addition of 100 µL fresh medium. A final concentration of 5% DMSO solution was added to positive control instead. After another 48 hrs incubation, all medium were discarded and cells were washed with PBS twice. 100 µL medium containing different DOX concentrations were added to treatment groups, respectively, as well as 20 µL MTS reagent. Scan was taken at 490 nm after 2 hrs incubation in CO2 incubator. This study confirms that there are no significant differences in readings between negative control and groups containing different DOX concentration up to 5 µg/mL.

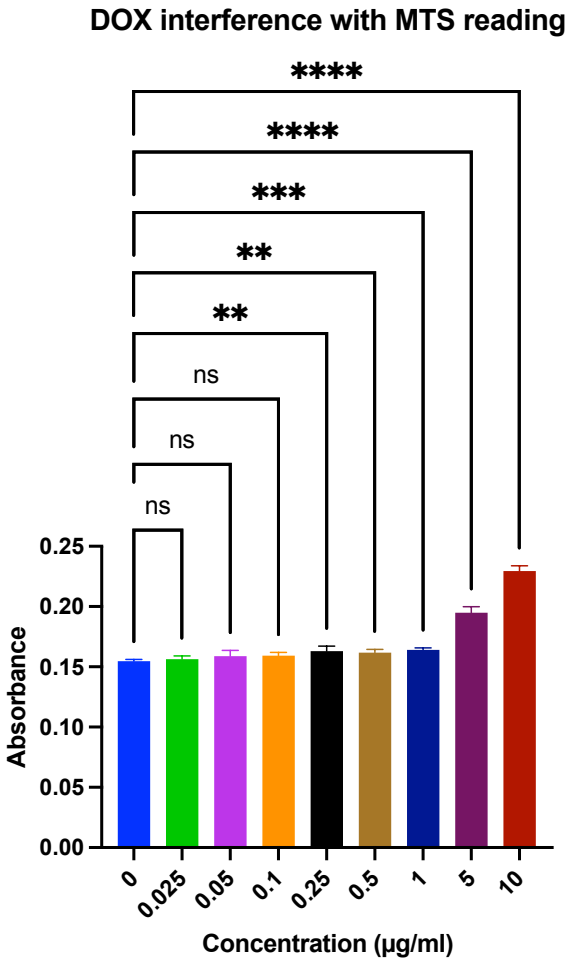
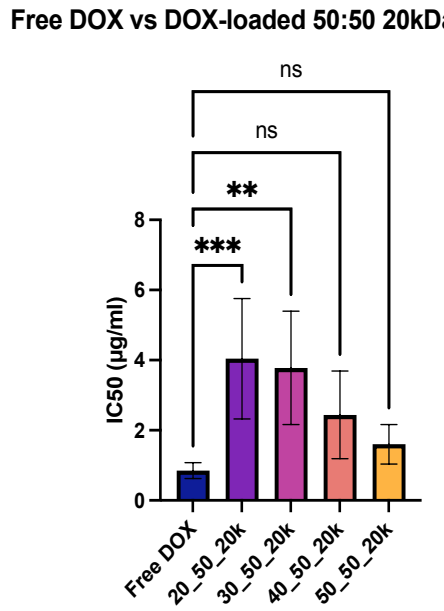


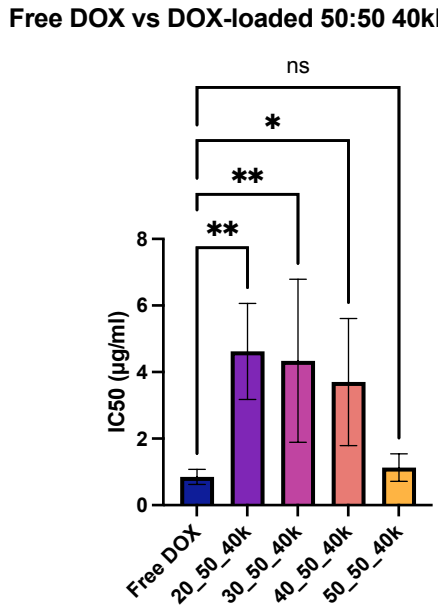
Figure B3. DOX interference with MTS assay in the absence of cells. This validation assay is done in a cell-free condition. 100 µL medium containing different DOX concentrations were added to treatment groups, respectively, as well as 20 µL MTS reagent. Scan was taken at 490 nm after 2 hrs incubation in CO2 incubator. This study confirms that there's no significant difference in readings between negative control group (0 µg/mL DOX) and treatment groups containing different DOX concentration up to 0.1 µg/mL.

Appendix C

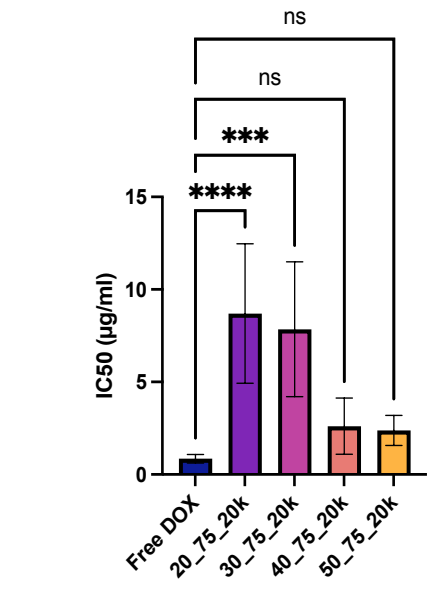
A.



B.



C.



D.

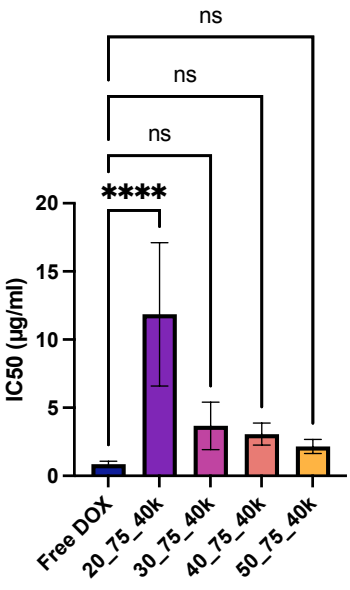
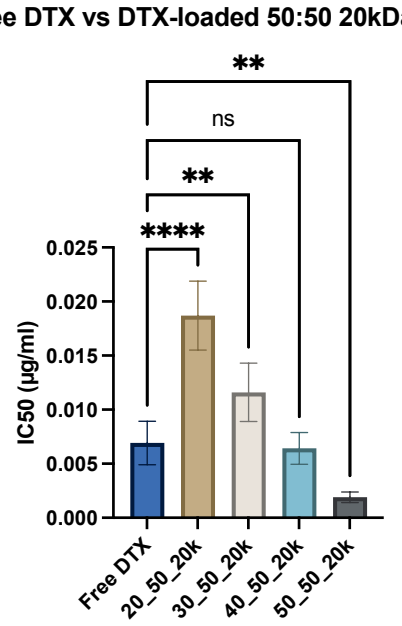


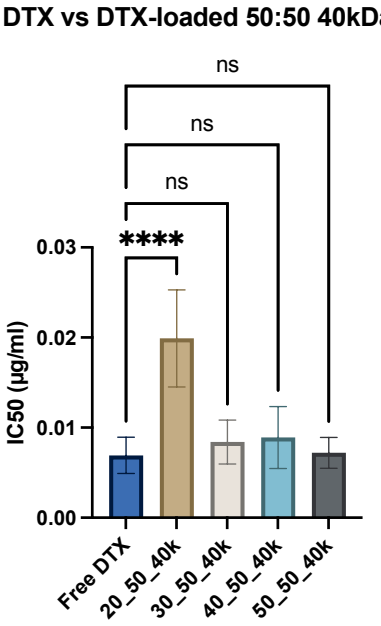
Figure C1. Comparison between free DOX IC50 and A) DOX-loaded 50:50 20kDa PLGA NP IC50, B) DOX-loaded 50:50 40kDa PLGA NP IC50, C) DOX-loaded 75:25 20kDa PLGA NP

IC50, D) DOX-loaded 75:25 40kDa PLGA NP IC50. One way ANOVA test was performed. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001. Data represent the mean ± SD (N=6).

A.



B.



C.

D.

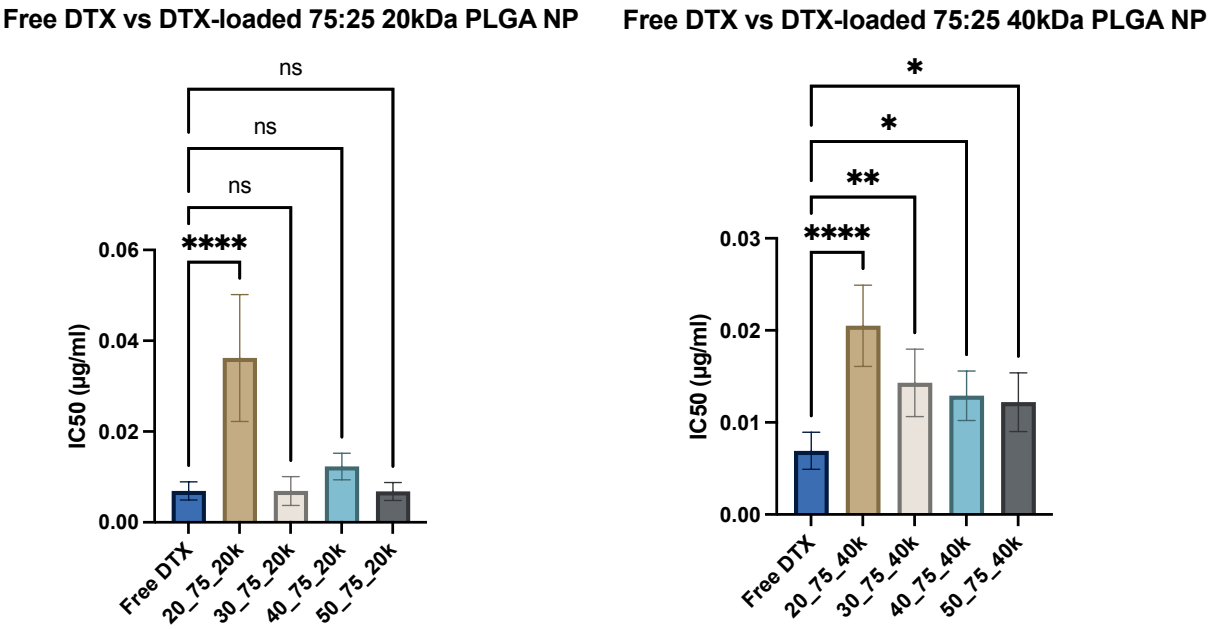
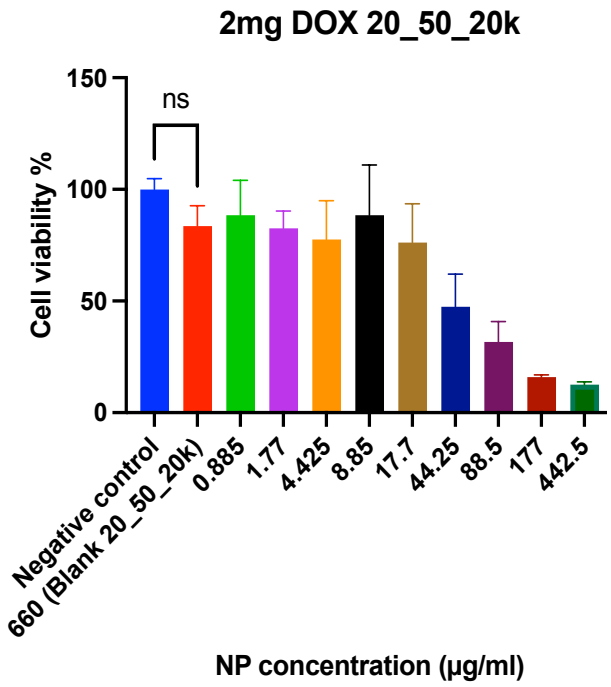


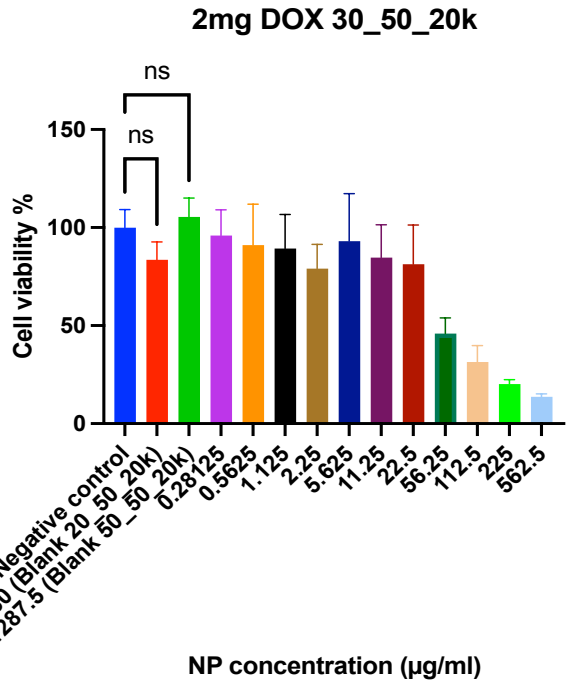
Figure C2. Comparison between free DTX IC50 and A) DTX-loaded 50:50 20kDa PLGA NP IC50, B) DTX-loaded 50:50 40kDa PLGA NP IC50, C) DTX-loaded 75:25 20kDa PLGA NP IC50, D) DTX-loaded 75:25 40kDa PLGA NP IC50. One way ANOVA test was performed. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001. Data represent the mean ± SD (N=6).

Appendix D

A.

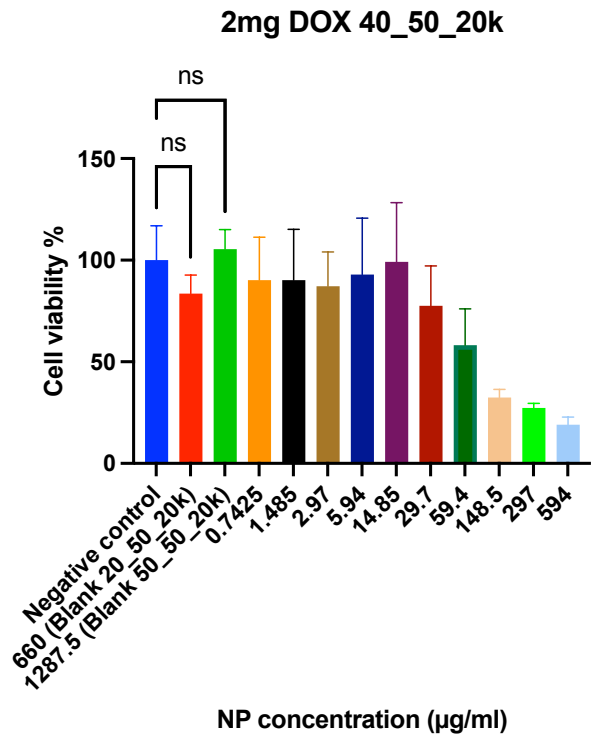


B.

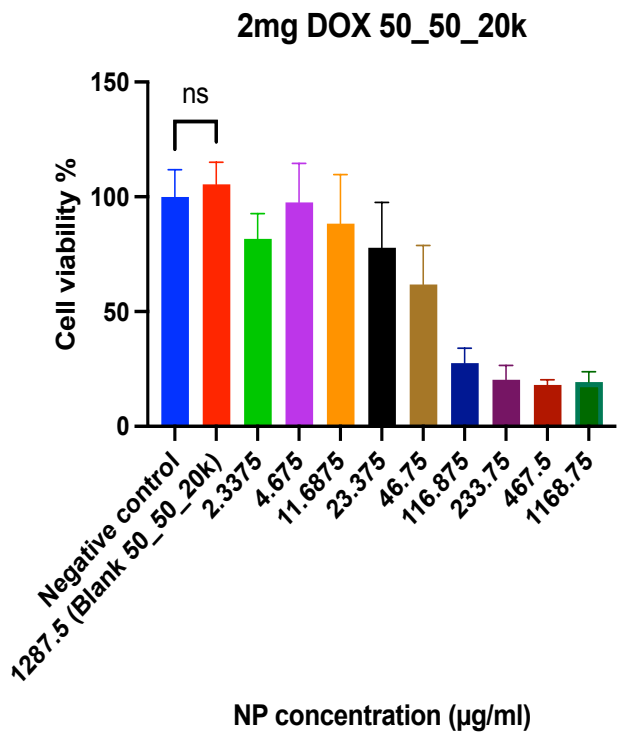


C.

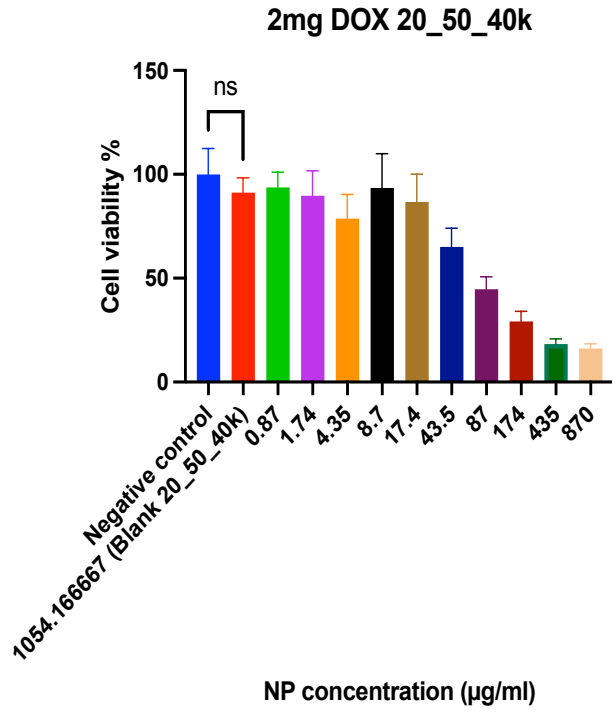
D.


2mg DOX 40_50_20k
ns
ns
150
100
50
0
Cell viability %
Negative control
660 (Blank 20_50_20k)
1287.5 (Blank 50_50_20k)
0.7425
1.485
2.97
5.94
14.85
29.7
59.4
148.5
297
594
NP concentration (µg/ml)

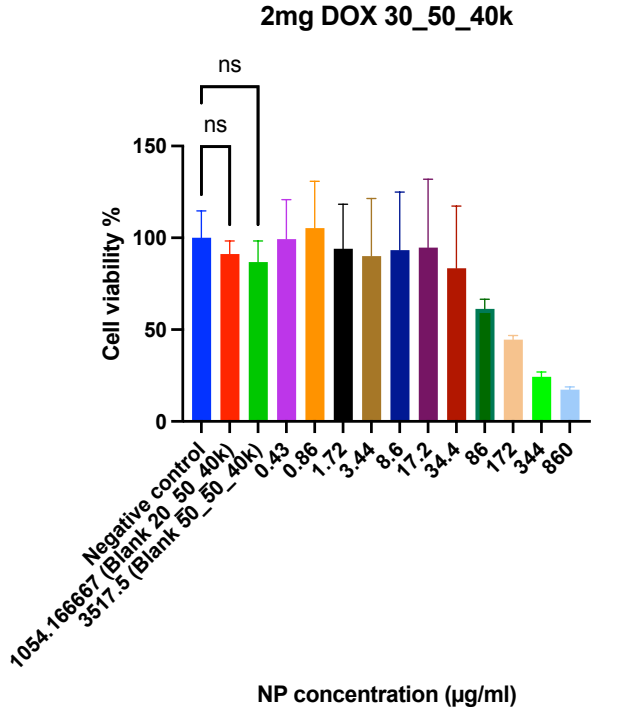
E.


2mg DOX 50_50_20k
ns
150
100
50
0
Cell viability %
Negative control
1287.5 (Blank 50_50_20k)
2.3375
4.675
11.6875
23.375
46.75
116.875
233.75
467.5
1168.75
NP concentration (µg/ml)

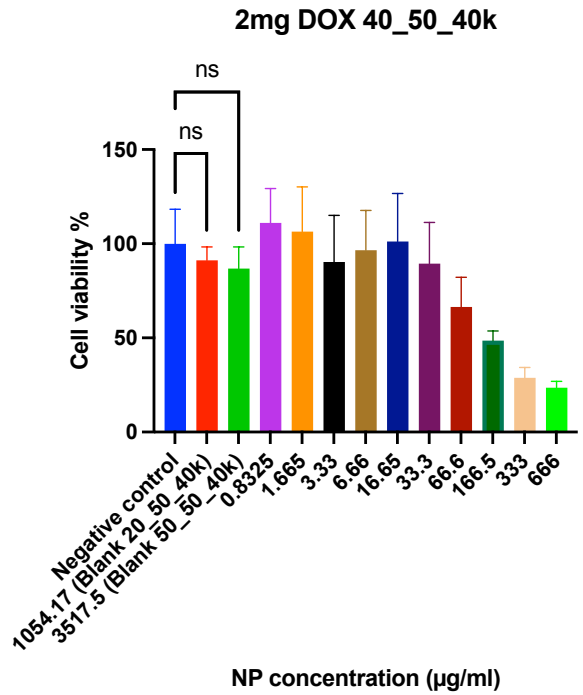
F.


2mg DOX 20_50_40k
ns
150
100
50
0
Cell viability %
Negative control
1054.166667 (Blank 20_50_40k)
0.87
1.74
4.35
8.7
17.4
43.5
87
174
435
870
NP concentration (µg/ml)

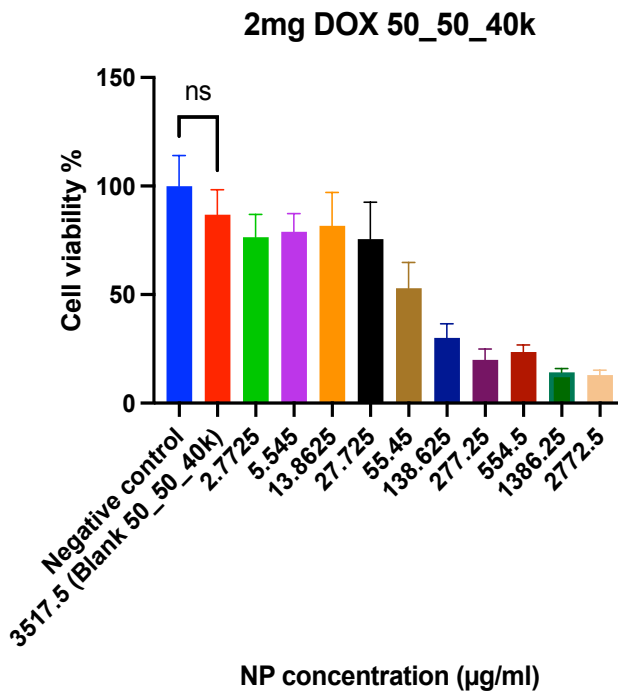
G.


2mg DOX 30_50_40k
ns
ns
150
100
50
0
Cell viability %
Negative control
1054.166667 (Blank 20_50_40k)
3517.5 (Blank 50_50_40k)
0.43
0.86
1.72
3.44
8.6
17.2
34.4
86
172
344
860
NP concentration (µg/ml)

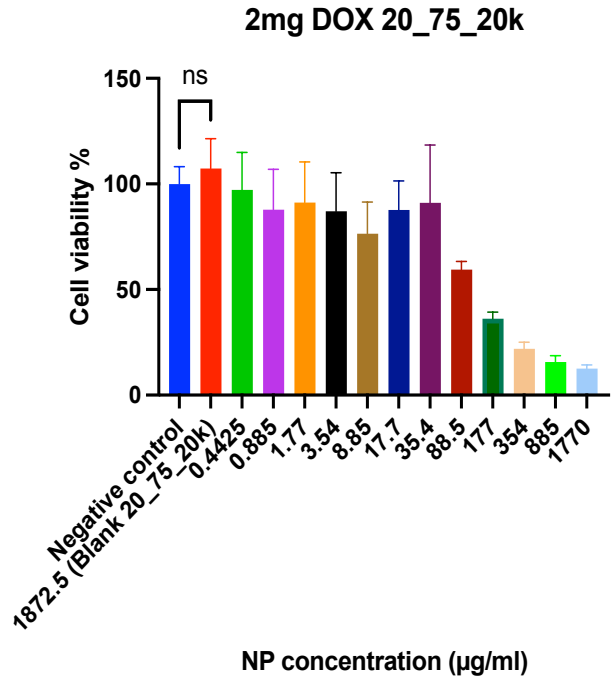
H.


2mg DOX 40_50_40k
ns
ns
Cell viability %
150
100
50
0
Negative control
1054.17 (Blank 20_50_40k)
3517.5 (Blank 50_50_40k)
0.8325
1.665
3.33
6.66
16.65
33.3
66.6
166.5
333
666
NP concentration (µg/ml)

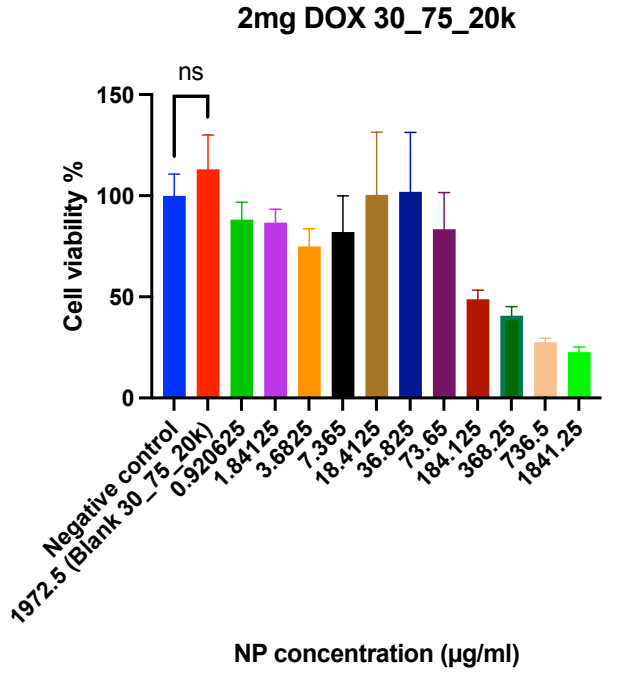
I.


2mg DOX 50_50_40k
ns
Cell viability %
150
100
50
0
Negative control
3517.5 (Blank 50_50_40k)
2.7725
5.545
13.8625
27.725
55.45
138.625
277.25
554.5
1386.25
2772.5
NP concentration (µg/ml)

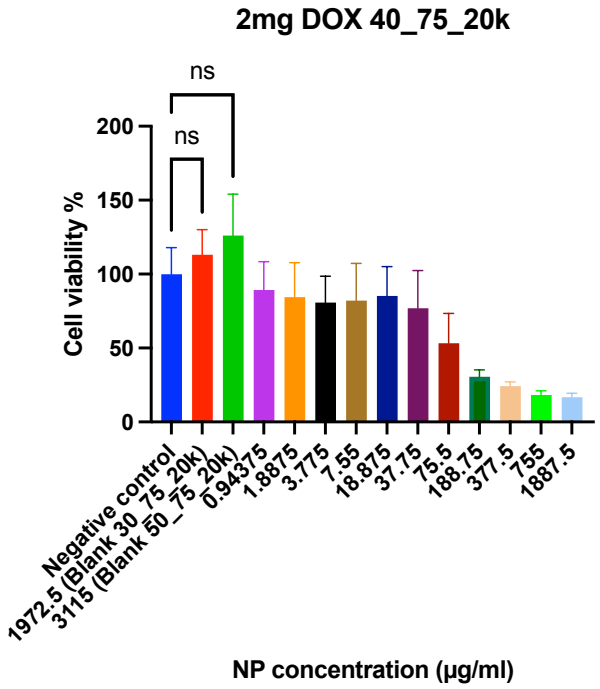
J.


2mg DOX 20_75_20k
ns
Cell viability %
150
100
50
0
Negative control
1872.5 (Blank 20_75_20k)
0.4425
0.885
1.77
3.54
8.85
17.7
35.4
88.5
177
354
885
1770
NP concentration (µg/ml)

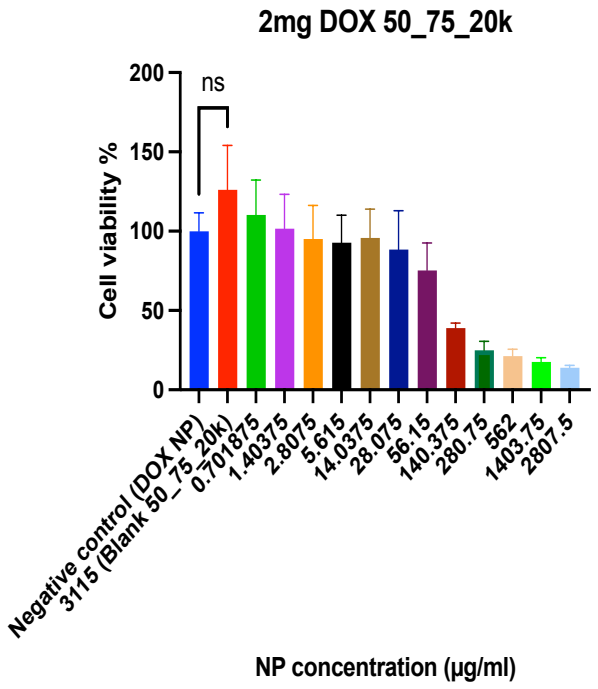
K.


2mg DOX 30_75_20k
ns
Cell viability %
150
100
50
0
Negative control
1972.5 (Blank 30_75_20k)
0.920625
1.84125
3.6825
7.365
18.4125
36.825
73.65
184.125
368.25
736.5
1841.25
NP concentration (µg/ml)

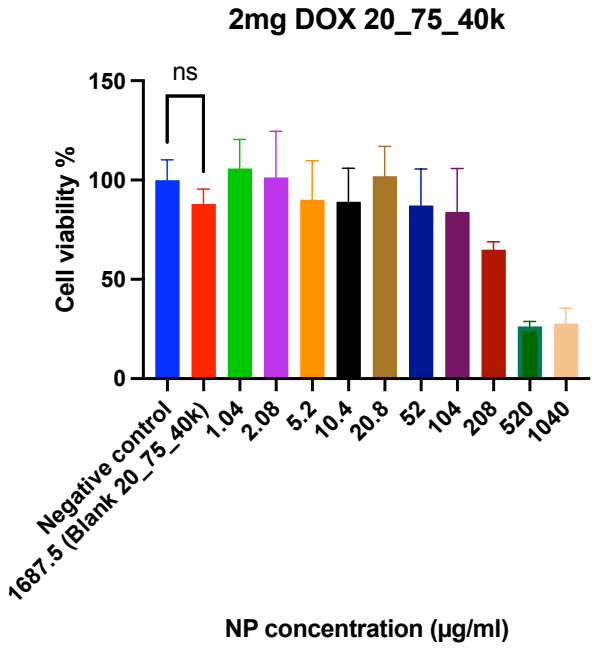
L.

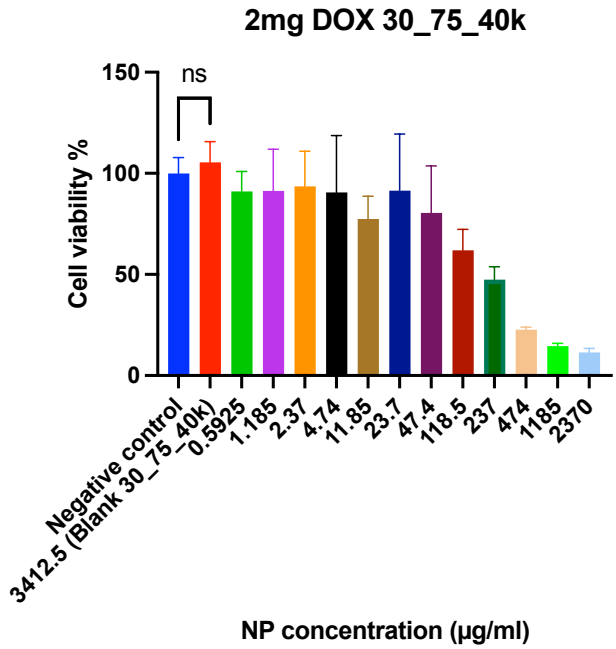

2mg DOX 40_75_20k
ns
ns
Cell viability %
200
150
100
50
0
Negative control
1972.5 (Blank 30_75_20k)
3115 (Blank 50_75_20k)
0.94375
1.8875
3.775
7.55
18.875
37.75
75.5
188.75
377.5
755
1887.5
NP concentration (µg/ml)

M.

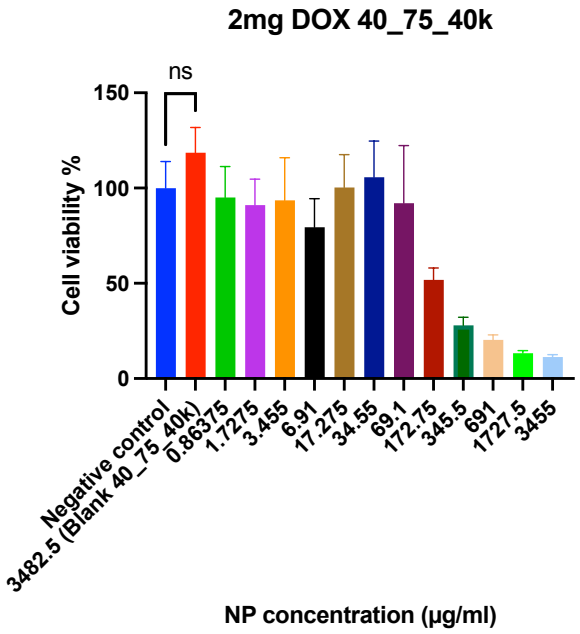

2mg DOX 50_75_20k
ns
Cell viability %
200
150
100
50
0
Negative control (DOX NP)
3115 (Blank 50_75_20k)
0.701875
1.40375
2.8075
5.615
14.0375
28.075
56.15
140.375
280.75
562
1403.75
2807.5
NP concentration (µg/ml)

N.


2mg DOX 20_75_40k
ns
Cell viability %
150
100
50
0
Negative control
1687.5 (Blank 20_75_40k)
1.04
2.08
5.2
10.4
20.8
52
104
208
520
1040
NP concentration (µg/ml)


2mg DOX 30_75_40k
ns
Cell viability %
150
100
50
0
Negative control
3412.5 (Blank 30_75_40k)
0.5925
1.185
2.37
4.74
11.85
23.7
47.4
118.5
237
474
1185
2370
NP concentration (µg/ml)

O.



P.

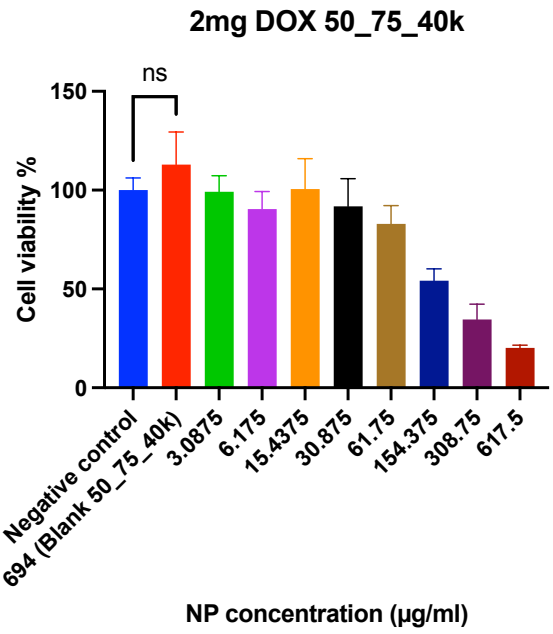
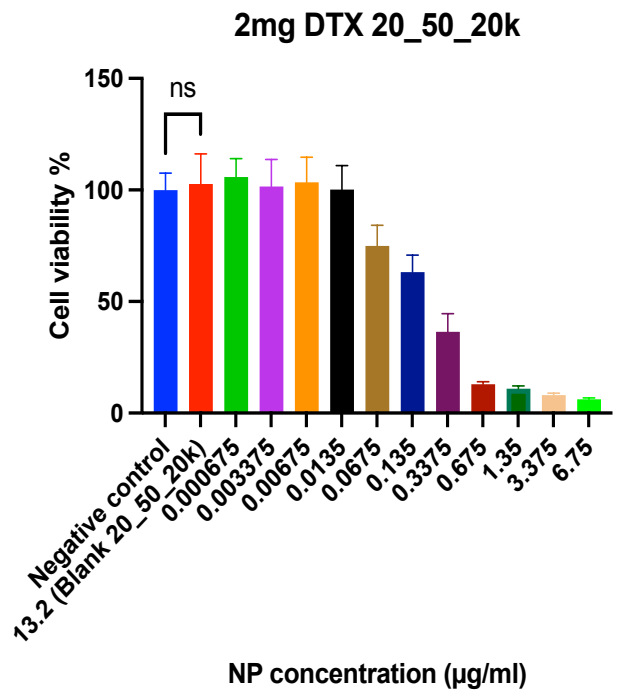


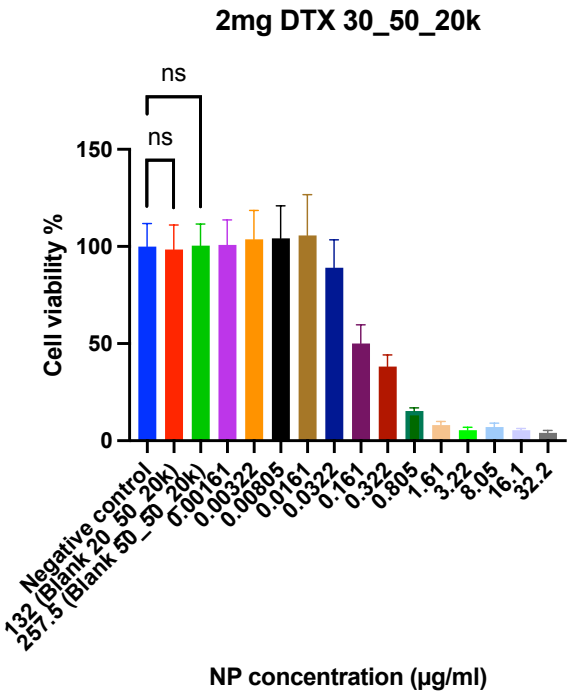
Figure D1. Blank NP and DOX NP cytotoxicity are compared to the negative control group which was not treated with any NP. Excess amount of blank NP was made to confirm that there was no cytotoxicity. One way ANOVA test was performed. Data represent the mean ± SD (N=6).

A.

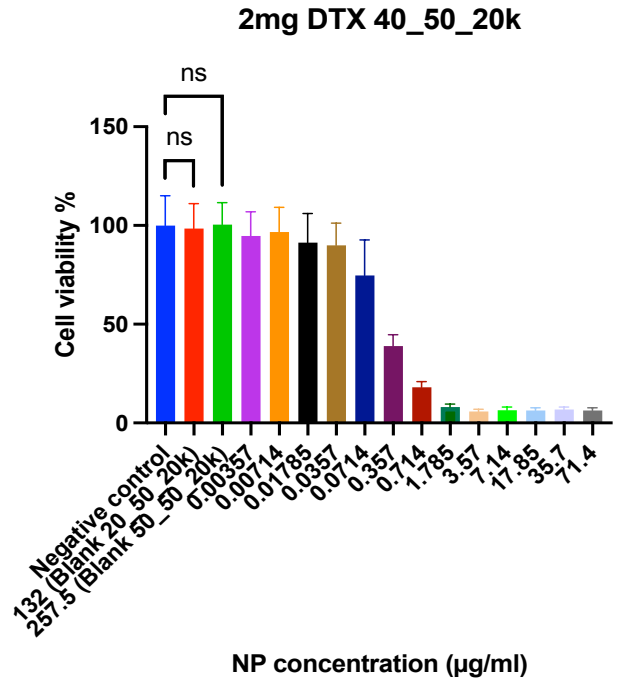
B.


2mg DTX 20_50_20k
Cell viability %
150
100
50
0
ns
Negative control
13.2 (Blank 20_50_20k)
0.000675
0.003375
0.00675
0.0135
0.0675
0.135
0.3375
0.675
1.35
3.375
6.75
NP concentration (µg/ml)

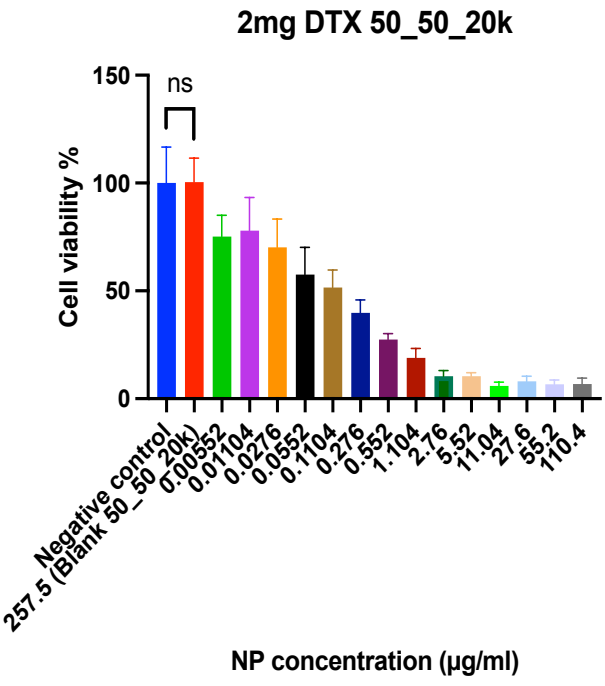
C.


2mg DTX 30_50_20k
Cell viability %
150
100
50
0
ns
ns
Negative control
132 (Blank 20_50_20k)
257.5 (Blank 50_50_20k)
0.00161
0.00322
0.00805
0.0161
0.0322
0.161
0.322
0.805
1.61
3.22
8.05
16.1
32.2
NP concentration (µg/ml)

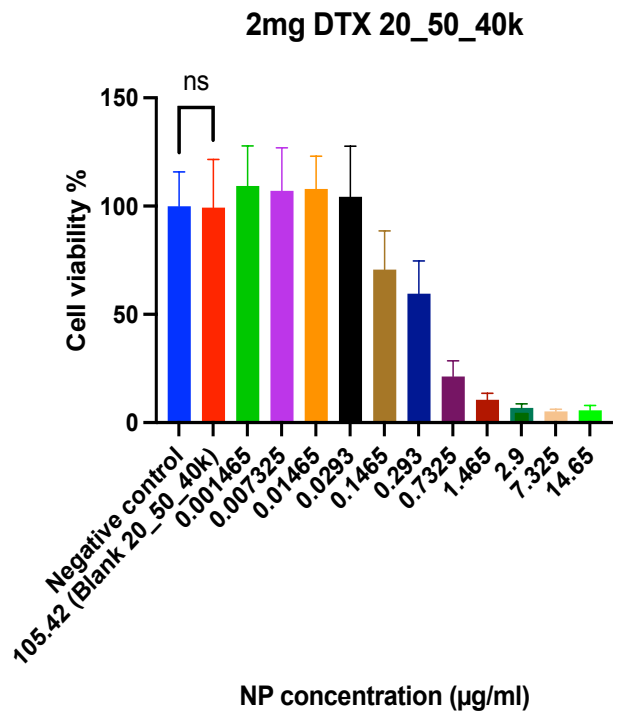
D.


2mg DTX 40_50_20k
Cell viability %
150
100
50
0
ns
ns
Negative control
132 (Blank 20_50_20k)
257.5 (Blank 50_50_20k)
0.00357
0.00714
0.01785
0.0357
0.0714
0.357
0.714
1.785
3.57
7.14
17.85
35.7
71.4
NP concentration (µg/ml)

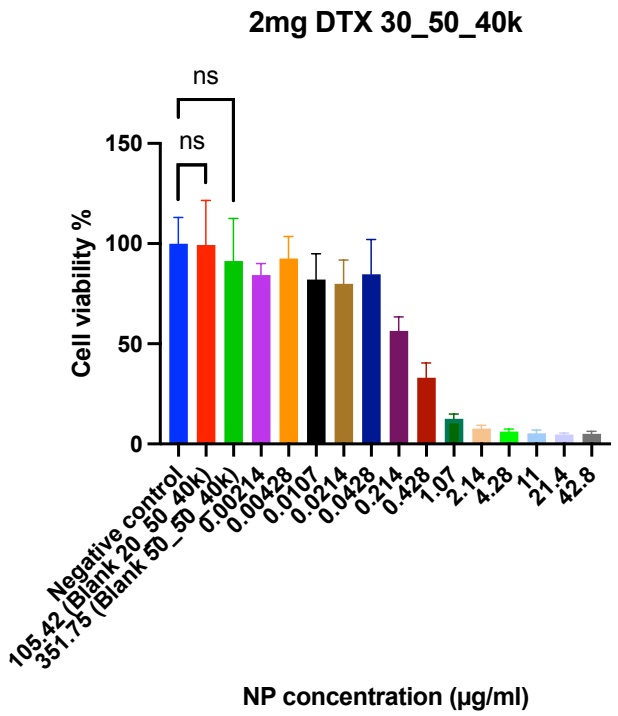
E.


2mg DTX 50_50_20k
Cell viability %
150
100
50
0
ns
Negative control
257.5 (Blank 50_50_20k)
0.00552
0.01104
0.0276
0.0552
0.1104
0.276
0.552
1.104
2.76
5.52
11.04
27.6
55.2
110.4
NP concentration (µg/ml)

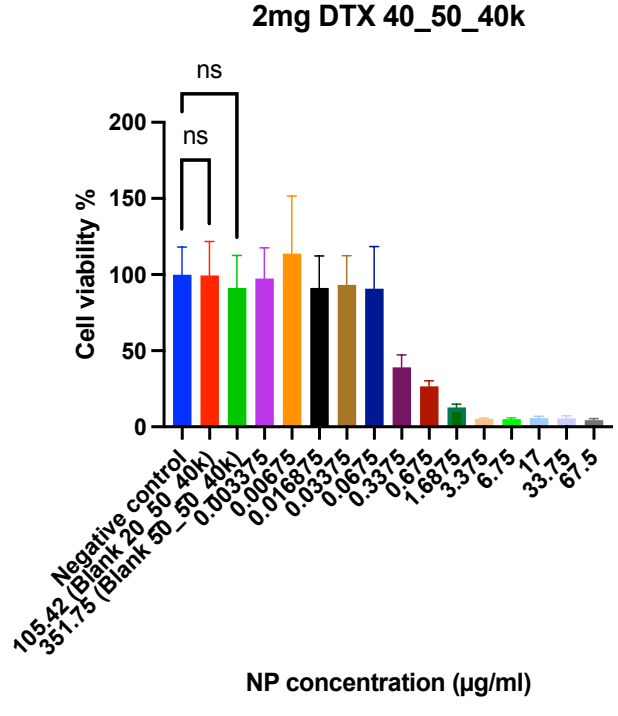
F.


2mg DTX 20_50_40k
ns
150
100
50
0
Cell viability %
Negative control
105.42 (Blank 20_50_40k)
0.001465
0.007325
0.01465
0.0293
0.1465
0.293
0.7325
1.465
2.9
7.325
14.65
NP concentration (µg/ml)

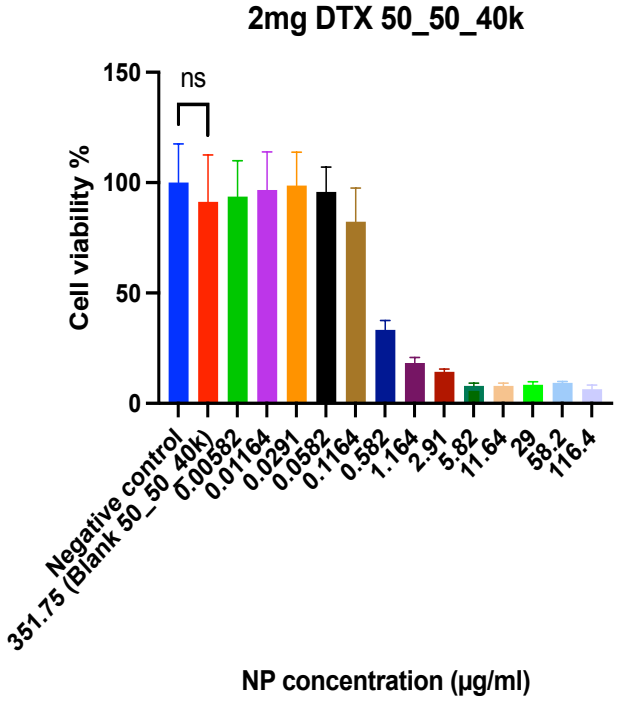
G.


2mg DTX 30_50_40k
ns
ns
150
100
50
0
Cell viability %
Negative control
105.42 (Blank 20_50_40k)
351.75 (Blank 50_50_40k)
0.00214
0.00428
0.0107
0.0214
0.0428
0.214
0.428
1.07
2.14
4.28
11
21.4
42.8
NP concentration (µg/ml)

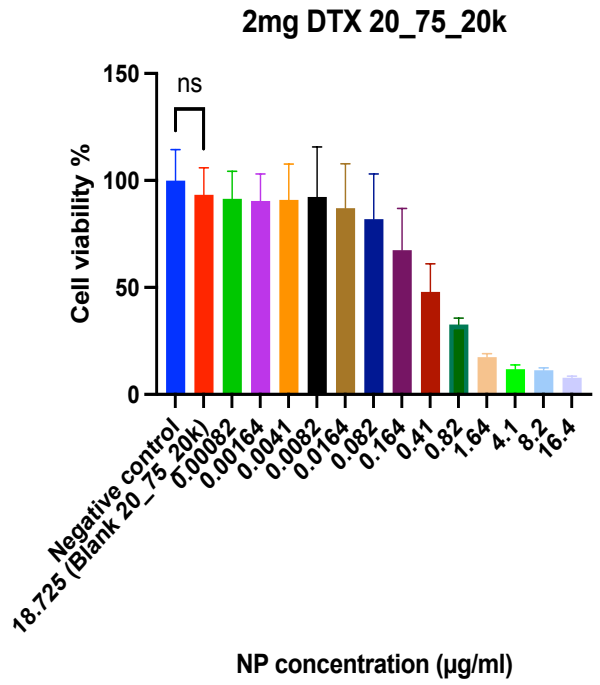
H.


2mg DTX 40_50_40k
ns
ns
200
150
100
50
0
Cell viability %
Negative control
105.42 (Blank 20_50_40k)
351.75 (Blank 50_50_40k)
0.003375
0.00675
0.016875
0.03375
0.0675
0.3375
0.675
1.6875
3.375
6.75
17
33.75
67.5
NP concentration (µg/ml)

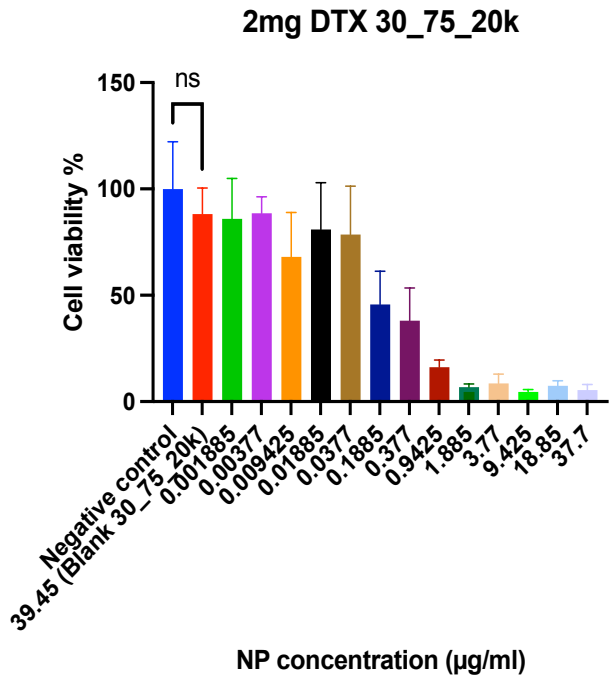
I.


2mg DTX 50_50_40k
ns
150
100
50
0
Cell viability %
Negative control
351.75 (Blank 50_50_40k)
0.00582
0.01164
0.0291
0.0582
0.1164
0.582
1.164
2.91
5.82
11.64
29
58.2
116.4
NP concentration (µg/ml)

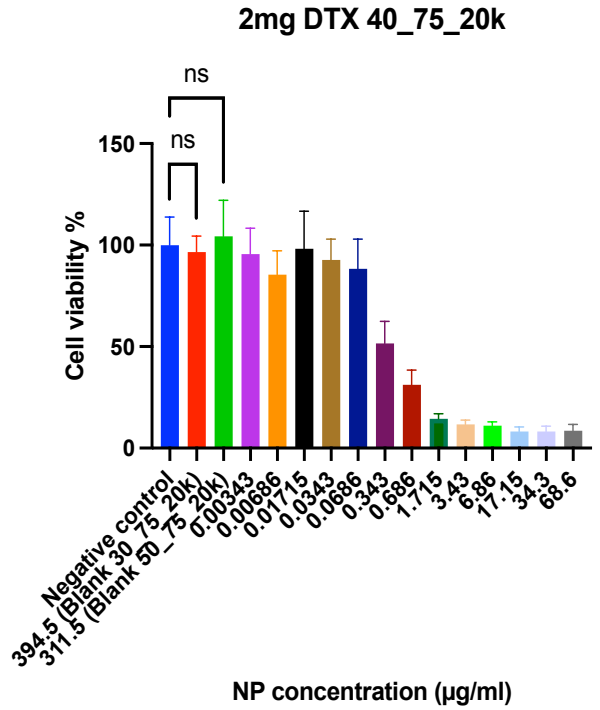
J.


2mg DTX 20_75_20k
ns
Cell viability %
150
100
50
0
Negative control
18.725 (Blank 20_75_20k)
0.00082
0.00164
0.0041
0.0082
0.0164
0.082
0.164
0.41
0.82
1.64
4.1
8.2
16.4
NP concentration (µg/ml)

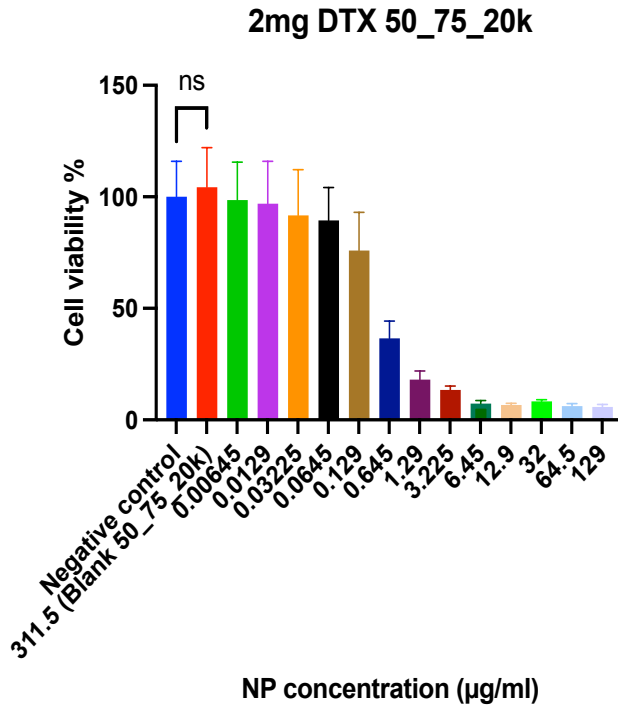
K.


2mg DTX 30_75_20k
ns
Cell viability %
150
100
50
0
Negative control
39.45 (Blank 30_75_20k)
0.001885
0.00377
0.009425
0.01885
0.0377
0.1885
0.377
0.9425
1.885
3.77
9.425
18.85
37.7
NP concentration (µg/ml)

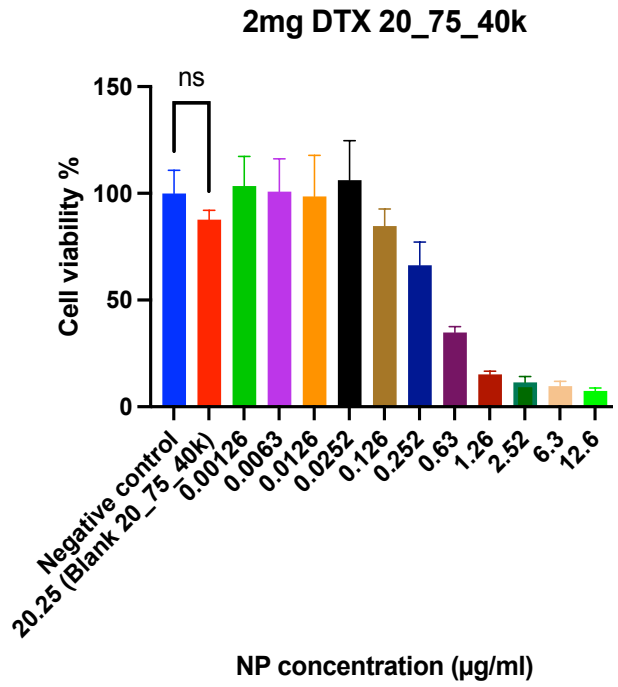
L.


2mg DTX 40_75_20k
ns
ns
Cell viability %
150
100
50
0
Negative control
394.5 (Blank 30_75_20k)
311.5 (Blank 50_75_20k)
0.00343
0.00686
0.01715
0.0343
0.0686
0.343
0.686
1.715
3.43
6.86
17.15
34.3
68.6
NP concentration (µg/ml)

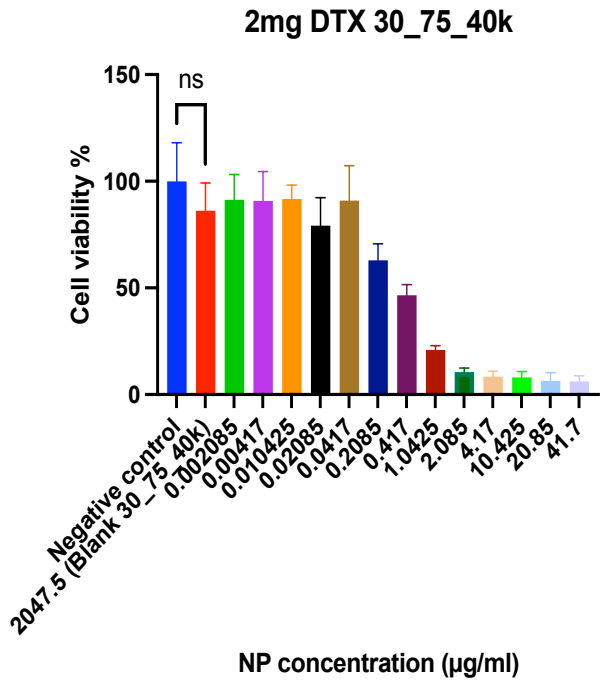
M.


2mg DTX 50_75_20k
ns
Cell viability %
150
100
50
0
Negative control
311.5 (Blank 50_75_20k)
0.00645
0.0129
0.03225
0.0645
0.129
0.645
1.29
3.225
6.45
12.9
32
64.5
129
NP concentration (µg/ml)

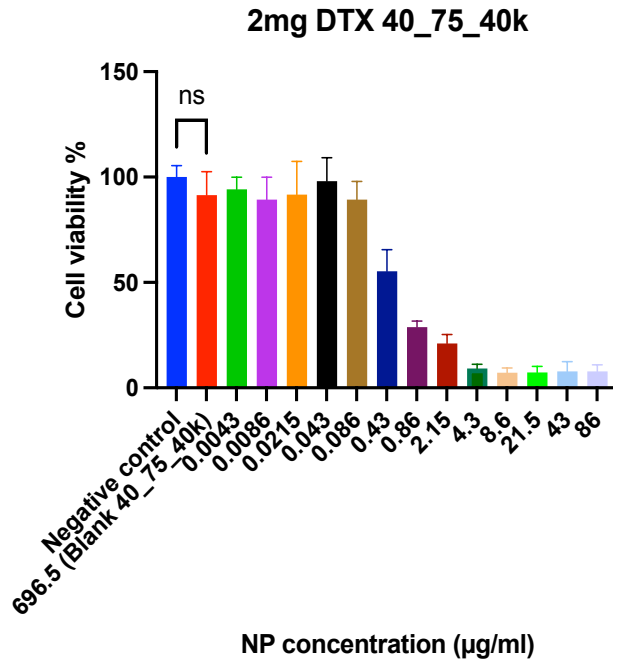
N.


2mg DTX 20_75_40k
ns
Cell viability %
150
100
50
0
Negative control
20.25 (Blank 20_75_40k)
0.00126
0.0063
0.0126
0.0252
0.126
0.252
0.63
1.26
2.52
6.3
12.6
NP concentration (µg/ml)

O.


2mg DTX 30_75_40k
ns
Cell viability %
150
100
50
0
Negative control
2047.5 (Blank 30_75_40k)
0.002085
0.00417
0.010425
0.02085
0.0417
0.2085
0.417
1.0425
2.085
4.17
10.425
20.85
41.7
NP concentration (µg/ml)

P.


2mg DTX 40_75_40k
ns
Cell viability %
150
100
50
0
Negative control
696.5 (Blank 40_75_40k)
0.0043
0.0086
0.0215
0.043
0.086
0.43
0.86
2.15
4.3
8.6
21.5
43
86
NP concentration (µg/ml)

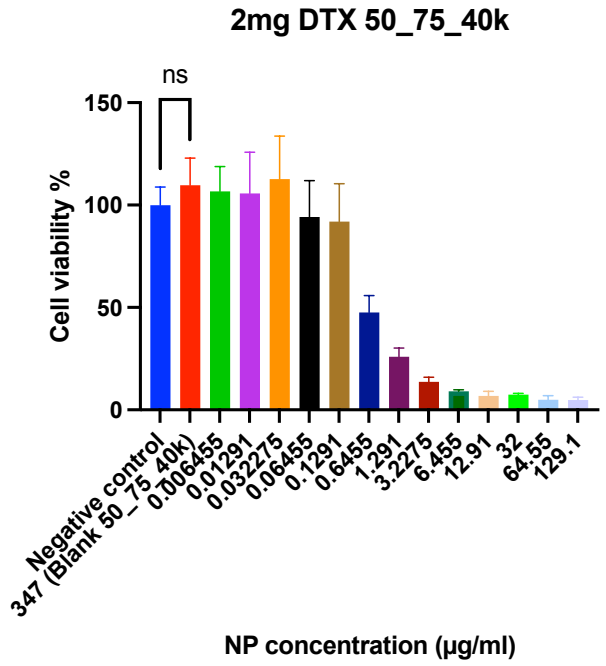

2mg DTX 50_75_40k
ns
Cell viability %
150
100
50
0
Negative control
347 (Blank 50_75_40k)
0.006455
0.01291
0.032275
0.06455
0.1291
0.6455
1.291
3.2275
6.455
12.91
32
64.55
129.1
NP concentration (µg/ml)

Figure D2. Blank NP and DTX NP cytotoxicity are compared to the negative control group which was not treated with any NP. Excess amount of blank NP was made to confirm that there was no cytotoxicity. One way ANOVA test was performed. Data represent the mean ± SD (N=6).