Feasibility and Application of Machine Learning Enabled Fast Screening of Poly-Beta- Amino-Esters for Cartilage Therapies

Stefano Perni
Cardiff University
Polina Prokopovich (prokopovichp@cf.ac.uk)
Cardiff University

Article

Keywords: Cartilage, PBAE, Machine Learning, Drug delivery, bagged MARS

Posted Date: March 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1422321/v1

Abstract

Despite the large prevalence of diseases affecting cartilage (with knee osteoarthritis affecting 16% of population globally), no curative treatments are available because of the limited capacity of drugs to localise in such tissues caused by the low vascularisation and electrostatic repulsion. While an effective delivery system is sought, the only option is using high drug doses that can lead to systemic side effects. We introduced poly beta amino esters (PBAEs) polymers to effectively deliver drugs into cartilage tissues. PBAEs are copolymer of amines and di-acrylates further end-capped with other amine, therefore encompassing a very large research space for the identification of optimal candidates.

In order to accelerate the screening of all possible PBAEs, the results of a small pool of polymers (n = 90) were used to train a variety of Machine learning (ML) methods using only polymers properties available in public libraries or estimated from the chemical structure. Bagged MARS returned the best performance and was used on the remaining (n = 3915) possible PBAEs resulting in the recognition of pivotal features; further refinements of such characteristics (n = 150) enabled the identification of a leading candidate predicted to improve drug uptake > 20 folds over conventional clinical treatment.

This work highlights the potential of ML to accelerate biomaterials development by efficiently extracting information from a limited experimental dataset thus allowing patients to benefit earlier from a new technology and at a lower price. Such roadmap could also be applied for other drug/materials development where optimisation would normally be approached through combinatorial chemistry.

1 Introduction

Biomaterial and drug design is regarded as a very resource (physical, economical and time) intensive operation [1] ; this process can be constructed into sequential stages (discovery, preclinical, clinical and pharmacovigilance) named Phase0 to Phase4. During Phase0, traditional bench experiments are carried out to identify optimal candidates that are screened through further developmental stages; while further clinical trials progressively assess toxicity, efficacy and long term safety (Phase1to Phase4) [2] . The overall development can take from a minimum of 5 up to 15 years with an estimated total development cost per approved drug of $2,168 million in 2018 [3]. However, the actual costs are generally a commercial confidential information and, therefore, such estimates may not fully capture the complete investments required [4] . The try-and-error approach to molecule development, particularly during the initial design and make phases of the design-make-test-analyse (DMTA) discovery cycle, is often is directed by human intuition, which is inherently biased and limited in knowledge, thus slowing drug development [5] . In such contest, the ability of data-driven in-silico prediction tools to model outcomes without the need to physically prepare candidates and run experiments would enable a fast throughput screening of candidate molecules and thus reducing both the time and monetary investments required to identify lead candidates [6–9] . This can be achieved by establishing correlations between certain properties of the molecules (inputs, also known as descriptors) and outcomes of interest using experimentally generated

data on a subset of relevant compounds; the established model would then be used to predict outcomes on the wider molecule search space [10] .

Machine learning (ML) based regression techniques are becoming wide spread in many areas of data analysis in the chemical [11,12] and pharmaceutical sector [13–16] and have recently been employed in drug development [17–19], diagnostic [20], treatment algorithm optimisation [21], drug repurposing [2,22] and material discovery [23,24]; however such applications are still quite limited despite been very promising [25,26]. Figure 1 depicts how ML could be deployed to accelerate the biomaterial development process. In spite of the flexibility of ML techniques, material design and optimisation involving numerous parameters are situations more likely to benefit from the development of machine learning predictive models.

Osteoarthritis (OA) is a thinning or loss of the cartilage layer covering the surfaces of joints reducing articular mobility also causing pain and inflammation. Although OA is not a life threating disease, it has a great impact on the quality of life and ability to perform regular activities of patients affected resulting in great burden to society and health care providers. Worldwide, 303.1 million of people live with hip or knee osteoarthritis [27]; furthermore, OA prevalence is expected to grow in consequence of the ageing population and overnutrition (two critical risk factors for OA). An effective treatment is still missing and current therapies (anti-inflammatory and analgesics) are only managing symptoms. This lack of therapeutic options is compounded by the inability of delivering the active molecules where is needed because of the obstacles posed by the low vascularisation and high electrostatic repulsion of cartilage tissues that limits the amount of drug effectively available to the targeted cells [28]. In order to achieve drug localisation, without a delivery system, high concentrations of drugs are used in the synovial fluid as mass transfer is governed by concentration differences (Fick's law) [29–31]. Such approach has some problematic drawbacks; firstly, it is a wasteful use of the drug as only a minimal amount is actually therapeutic, with consequences on treatment acquisition costs. Secondly, drug washout lead to systemic exposure with possible side effects, as in case of steroids [32].

Different drug delivery systems have been developed for the localisation of drugs in cartilage in the attempt to overcome such barriers; Poly-beta-amino-ester (PBAEs) [33,34] and avidin [29] are two examples of these delivery systems. While no particular optimisation of the delivery system based on avidin performance is feasible as this a well-defined protein; there are, instead, essentially ∞^2 possible PBAEs as these are copolymers of an amine and a di-acrylate [35]. Moreover, when PBAEs end-capping is also considered, the possible combinations rise to ∞^3. In light of the performance of PBAE as cartilage drug delivery system being extremely dependent on the polymer backbone; ML algorithms predicting the efficacy of the drug delivery in cartilage by the delivery system from the polymer's constituents' properties would provide a high throughput screening for the optimisation of the PBAE driven cartilage drug localisation technology reducing the cost and time to select the most promising candidate. We have previously demonstrated how the uptake of dexamethasone (DEX) (a drug routinely administered in clinics through intra-articular injections to reduce OA symptoms) in cartilage tissue, through a poly-beta-amino-ester drug delivery system, could be modelled using partial least square regression [33,34]. The

inputs of this model are the physical properties of the polymers and co-polymeric units (di-acrylate and amine) along with some experimentally obtained parameters such as the diffusion coefficient of the polymer through cartilage, the drug loading in the delivery system and the molecular weights (Mw and Mn) of the polymer chain [34]. Despite the ability of predicting uptake, this model, in order to make predictions on new candidates, still required inputs generated by experiments (such as Mw, Mn and diffusion coefficient) thus not fully able to completely substitute lab-based work. Through this previous work, we identified a polymer (current lead candidate obtained from screening the combination of 3 acrylates and 15 amines) that increased DEX uptake in cartilage about 8 times compared to the clinical formulation [34]. With the purpose of accelerating the optimisation of the PBAE structure for the cartilage delivery system through a systematic screening of large library of both acrylates and amines, we hypothesised that machine learning algorithms utilising only predictors available in public libraries or calculated from the compound structure, namely the physico-chemical properties of the PBAE components, could be employed to fully predict the performance of the delivery system without the need for any experimentally originated data. Drug utilised uptake data experimentally obtained from a subset of a large polymer library were to train and optimise 25 machine learning models (e.g. Random Forests, kNN, SVM, neural network and MARS) and investigated their predictive performance to identify the most accurate algorithm. This model was then employed to screen the PBAEs research space and key features in the amine and acrylate structure recognised, further elucidating correlations between PBAE structural properties and drug uptake. A further round of ML predictions was conducted on variations of the previously selected core structures to refine and improve efficacy. The most promising candidate identified had 3 folds expected efficacy improvement over the previous best performing candidate.

2 Results

2.1 Machine learning model selection

Amine 1 to 20, acrylates A to F and end-capping e-1 and e-2 were used to generate the library of PBAE-DEX used for the experimental determination of DEX uptake in cartilage; in total 15*6*2 = 180 unique PBAE were synthesised, doubling the size of the experimentally tested PBAE. After random splitting, the train set included 70 PBAEs, while the remaining 20 PBAEs constituted the test set. As the ultimate purpose of modelling is being able to estimate outcomes (in our work the uptake of DEX in cartilage) on previously unseen predictors, a split of the initial dataset into train and test set was implemented to be able to identify the model with the greatest predicting ability that is not necessarily the one that return the most accurate fit of the data used to calculate the model parameters (i.e. regression coefficients) [36,37]. For the same reason, data split in training and test set was stratified based on PBAEs thus experimental data of DEX uptake for different exposure duration and related to PBAE with different end-capping all belonged to one set only. The 25–75% also is in the typical range to provide sufficient data points for both model parameters estimation (training set) and testing [38–41]. Therefore, it was expected that all models performed better on the training set than on the test set obtained from the experimentally test PBAE library (Fig. 2).

Bagged multivariate adaptive regression splines (bagged MARS) returned the lowest RMSE on the test set (0.072). Random Forest had the lowest RMSE on the training set (0.036) but the second lowest on the test set (0.073). However, regressions based on decision trees/random forests do not allow for extrapolation of the measured outcome beyond the training set and such would limit the possibility of identifying PBAE performing better the experimentally observed optimal candidate. Linear regression (forward, backward or stepwise) had the highest RMSE on the training datasets, 0.128, 0.128 and 0.124, respectively, on training and test dataset. The difference in model performance between train and test set depended on the algorithm used, for example elastic regulation had RMSE of 0.080 and 0.081 for train and test set respectively, while Bayesian addictive regression trees returned RMSE of 0.043 and 0.149 on train and test set, respectively. The small difference between the RMSE on train and test set observed for the elastic regulation model is a consequence of the penalties assigned to predictors in the algorithm that reduce the risk of overfitting [37,42]. Moreover, boosting and bagging improved model performance (Fig. 2), for example RMSE of bagged MARS lower than MARS, random forests better than decision tree. This was expected as such approaches have been developed to improve on model performance [37,42]. Bagging is the process of resampling from the same data set to generate numerous new datasets then used to fit the model, this bootstrapping reduces overfitting and model variability [37,42]; on the other hand, boosting employs weak predictors to improve on the predictions of other predictors [43].

The optimisation of the bagged MARS model hyper-parameters showed that with increasing number of bagged samples mean RMSE during cross-validation decreased; averaging 75 resamples gave the lowest RMSE (Fig. 3a) while the number of pruned parameters increased model performance monotonically, but RMSE marginally decreased with the combination of more than 10 (Fig. 3b). Moreover, performance of bagged MARS improved when the degree of interaction between parameters increased from 1 to 2 (Fig. 3b). The optimal bagged MARS model was made of a combination of 75 MARS models with a median number of predictors of 9 and a median number of terms of 15.

DEX uptake predicted by the optimised bagged MARS model against the actual data for the test data set (Fig. 4a) revealed a general good agreement between prediction and actual data regardless of the PBAE end-capping agent while the residual distribution exhibited a gaussian distribution (Fig. 4b). Similar patterns were observed when the model was applied on the train set (Fig. 4c and d); however, the residuals were smaller resulting in a narrower distribution. Modelled uptake curves of DEX in cartilages with PBAE in the test set demonstrated a general good fit of the experimental data (Fig. 4e).

The variable with the greatest importance in the bagged MARS model was ZStericQuad3D of the amine component, followed by the complexity of the amine component and the Henry's law coefficient of the PBAE repeated unit, the variable with the lowest importance was the molecular weight (MW) of the acrylate component (Fig. 5a). In order to gain insights on the relations between the chemical and topological properties of the PBAE and the efficacy in localising DEX in cartilage, the specific dependence of the DEX uptake on the individual predictors was analysed on through the partial dependency plot (PDP).

These plots represent the predicted outcomes against a single varying input variable while maintaining the remaining constant at their mean values. PDP revealed ZStericQuad3D returned a maximum DEX uptake at ~ 0.83; while complexity of the amine decreased DEX uptake for values up to 50, for greater amine complexity predicted DEX uptake increased monotonically but was lower than the maximum (complexity = 0) for the maximum amine complexity in the library tested (Fig. 5b). As the models were trained on transformed values the relations between variables and drug uptake does not appear linear on the back-transformed predictions.

2.2 PBAE structure optimisation

The optimised bagged MARS model was applied on the remaining PBAE search space constituted by 3915 un-synthesised polymers to predict DEX uptake in cartilage after 10 min of exposure to PBAE-DEX when end-capped with e-1 or e-2. The results of this round1 screening identified 3192 PBAEs, regardless of the end-capping (end-capping agent e-1 returning predominantly higher drug uptake than e-2 on the same PBAE backbone), with an expected DEX uptake greater than the commercial formulation. Furthermore, 11 polymers with a predicted uptake greater than the previous leading candidate, which returned a drug uptake about 8 times that of DEX commercial formulation, were identified through the model. These PBAE clustered very closed according to the dendrogram determined using the chemico-physical properties of the polymers and were made mainly by acrylate AAA (Phenylmethanediol diacrylate) or XX (1,4-Phenylene diacrylate) and amine 69 (2-Amino-5-(cyclopropyl)pyrazine) or 70 (2-Amino-6-propylpyrazine). PBAE XX-69 was predicted to exhibit the greatest uptake among the full PBAE library tested, about 13 folds greater than the commercial formulation (Fig.6).

1,4-phenylene diacrylate and (acrylate XX) and phenyl-methanediol diacrylate (acrylate (AAA) are the only acrylates tested exhibiting a benzene group where the electron of the oxygen atoms forming the di-acrylate groups can resonate reducing the impact of the electrostatic repulsion between some areas of the PBAE backbone and GAG constituents of cartilage. Similarly, the presence of pyrazine in the amine constituent can increase the availability of the electron pair in the nitrogen resulting in higher positive charge. Further acrylates exhibiting at least a benzene group in proximity of the acrylate moiety along with amines with a pyrazine in their structure (Figure S5) were screened in Round2. 17 PBAE had an estimated DEX uptake greater than XX-69 (best performer in round1); the presence of a further tertiary amine bound to the pyrazine ring resulted in greater DEX uptake in cartilage; moreover, two benzine groups (Bisacrylic acid oxybis(4,1-phenylene) ester) improved on the drug delivery (Fig. 7). The most effective PBAE (DDD-114) identified in round2 had a predicted DEX uptake about 21 time greater than the commercial formulation.

3 Discussion

The key to accurate predictions through mathematical models is the size of the data set used for the estimation of the model parameters [44]. As our previous work hinted to the possibility of modelling cartilage drug uptake achieved by PBAEs conjugated to DEX [34], the machine learning models in this work

were trained using a dataset [34] doubled in size with further polymers to reach a sufficient level of confidence in the model estimates. The work presented here considers only two end-capping agents treated as a categorical variable; the actual properties of the compounds were not considered as the number of molecules did not allow to capture such possible parameters.

Majority of research dedicated to implement ML in drug discovery/chemistry employs a very narrow range of potential models, even just one [44–46], without a clear rationale for the selection of the algorithms included in the pool assessed [5,18,47–50]. Here instead, we purposely screened a large number of potential algorithms based on different approaches (e.g. decision tree, linear regression, SVM and neural network) in order to maximise the strength and transferability of the results while, simultaneously, increase the likelihood of identify a satisfactory predictive model.

MARS are an extension of linear models that can account for nonlinearities between input and output values through the use of hinge functions and interactions between variables combining flexibility and interpretability of results [37,42]. The overall regression model “goodness of fit” depends on hyperparameters such as the number of pruned parameters and the degree of interaction between predictors. Bagged MARS is an ensemble of MARS constructed on a randomly generated bootstrapped set of data. Although it was expected that aggregating further resamples would improve model predictive performance, no more than 75 resamples were implemented in this work as the reduction in RMSE from 50 to 75 resamples was already minimal and a further increase of the resamples would also impact computational time.

The electrostatic interactions between positively charged PBAE and cartilage tissue components (predominantly the highly negatively charged glycosaminoglycans - GAGs) are the key mechanism of action of the delivery system under the presented investigation. The ranking of the PBAE properties variables showing quadrupole on the Z axis of the amine component as the key parameter demonstrated by the analysis of variable importance is in agreement with the mechanisms of action and it was also found to be one the key parameters when PLS regression was carried out using not only chemico-physical properties but also experimentally determined characteristics (diffusion coefficient, zeta potential and molecular weight of the polymer) [34].

The optimal components identified here are structurally very different from those found as optimal copolymers for PBAE application in DNA vector [35,51] and a direct consequence of the different mechanisms involved in the technology (DNA binding and cell membrane penetration vs. electrostatic attraction toward negatively charged GAG chains in cartilage).

The application of ML to PBAE structure optimisation for drug delivery in cartilage presented in this work can also potentially act as blueprint for the optimisation of other applications of PBAE such as drug releasing degradable coatings [52], non-viral DNA vectors for gene therapy [35] and mRNA vaccines [51] fast-tracking products to patients where, to date, only a lab based combinatorial chemistry approach to optimisation has been undertaken [53]. The expected reduction in the time required to screen numerous

polymers will also be coupled with monetary saving in the drug development costs with clear benefits not only to patients but also to health care providers.

We demonstrated an ML guided drug design optimisation approach that accurately predicts the relation between structure/property and outcome requiring only 2% of the compositional space (90 out of at least 3915 copolymers) to be experimentally explored. Our worked led to the discovery of several PBAEs expected to result in a higher drug uptake than those of previously reported candidates. Moreover, the trends uncovered between properties and efficacy of the polymers along with the nonintuitive optimal design elements of PBAE for cartilage delivery identified in this study, such as the presence pyrazine in the amine constituent (likely related to the increased hydrogenation of the nitrogen atom) are also critical in the search for next-generation polymer driven cartilage delivery systems.

4 Methods

PBAE are denoted throughout the text with a code containing letters referring to the diacrylate (Figure S1) and numbers (Figure S2) referring to the amine; for example, A5 is the polymer made from 1,4-Butanediol diacrylate and 3-(dimethylamino)propylamine. The polymer backbone code is followed by e1 for PBAE end-capped with ethylene-diamine and e2 for PBAE end-capped with diethylene-triamine.

4.1 Data analysis

All models were fitted through R [54] and all other necessary packages necessary to perform regression with the “caret” package [55].

Manhattan distance between PBAEs and complete distance between clusters were used for generating dendrograms.

4.1.1 Datasets and Descriptors

Two PBAE uptake datasets were used to develop predictions, the publicly available set[34] was expanded with a purposely obtained new set collected after the inclusion of further acrylate monomers in the library.

Drug uptake predictions were performed utilizing physical and chemical parameters of amine and acrylate components of each PBAE obtained from PubChem library (Mw, logP, tPSA, Complexity, Heavy Atom Count, Volume 3D, X_Steric Quadrupole 3D, Y_Steric Quadrupole 3D, Z_Steric Quadrupole 3D); along with parameters related to the repeated polymeric unit (amine + acrylate) calculated through ChemDraw. The later included boiling point, melting point, critical volume and pressure, Gibb's free energy, logP (partition-coefficient between two immiscible phases at equilibrium which is proportional to hydrophobicity), solubility (logS), pKa, molar refractivity (CMR), heat of formation and the topological polar surface area (tPSA), which represent the total area of all polar atoms (mainly oxygen and nitrogen) including their affixed hydrogen atoms.

K^{th} nearest neighbour imputation was employed to handle missing data [42].

4.1.2 Machine learning algorithms training and predictions

Outcome data were transformed ($1/ y^4$) to achieve a distribution of the drug uptake closer to a gaussian profile; moreover, input values for possible predictive variables were centred and scaled using mean and standard deviation.

A random split of the PBAEs into training (75%) and test (25%) datasets was applied. Weights to each point were assigned proportionally based on the distance from the median. Classic Machine Learning methods, such as Bernoulli Naive Bayes, Elastic regularisation, K^{th} nearest neighbour (kNN), generalised addictive models (GAM), Decision Tree, Random Forests, Neural Networks and Support Vector Machine (SVM) were employed to establish correlations between predictors and drug uptake. Tuning and hyper parameters search for each model were conducted through 10-fold cross validation repeated 3 times on the training dataset; final model selection was based on minimisation of RMSE. The same training and test data set were employed for all models tested.

The best performing predictive model was used to estimate the drug uptake of the PBAE not previously experimentally tested in a two-steps approach. During the first round, amine and acrylates exhibiting a variety of structural features and moieties was employed to recognise critical patterns. In round2, variations of the pivotal characteristics observed in round1were explored to further refine the optimal candidate.

Declarations

Declaration of competing interest

The authors are named inventors on patents related to the use of PBAE as drug delivery systems.

Acknowledgments

The work has been supported by Wellcome Trust Pathfinder Fund.

Authors contribution

SP: Conceptualization, Methodology, Software, Investigation, Formal analysis, Data Curation, Visualization, Writing - Original Draft.

PP: Validation, Methodology, Funding acquisition, Resources, Writing - Review & Editing, Project administration.

Availability of Data

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

References

1. Gupta, R. *et al.* Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol Divers **25**, 1315–1360, doi:10.1007/s11030-021-10217-3 (2021).
2. Réda, C., Kaufmann, E. & Delahaye-Duriez, A. Machine learning applications in drug development. Computational and Structural Biotechnology Journal **18**, 241–252, doi:https://doi.org/10.1016/j.csbj.2019.12.006 (2020).
3. Solutions., D. C. f. H. (2018).
4. Morgan, S., Grootendorst, P., Lexchin, J., Cunningham, C. & Greyson, D. The cost of drug development: a systematic review. Health Policy **100**, 4–17, doi:10.1016/j.healthpol.2010.12.002 (2011).
5. Reis, M. *et al.* Machine-Learning-Guided Discovery of 19F MRI Agents Enabled by Automated Copolymer Synthesis. Journal of the American Chemical Society **143**, 17677–17689, doi:10.1021/jacs.1c08181 (2021).
6. Ekins, S. *et al.* Exploiting machine learning for end-to-end drug discovery and development. Nature Materials **18**, 435–441, doi:10.1038/s41563-019-0338-z (2019).
7. Struble, T. J. *et al.* Current and Future Roles of Artificial Intelligence in Medicinal Chemistry Synthesis. Journal of Medicinal Chemistry **63**, 8667–8682, doi:10.1021/acs.jmedchem.9b02120 (2020).
8. Paul, D. *et al.* Artificial intelligence in drug discovery and development. Drug discovery today **26**, 80–93, doi:10.1016/j.drudis.2020.10.010 (2021).
9. Kimber, T. B., Chen, Y. & Volkamer, A. Deep Learning in Virtual Screening: Recent Applications and Developments. Int J Mol Sci **22**, 4435, doi:10.3390/ijms22094435 (2021).
10. Moosavi, S. M., Jablonka, K. M. & Smit, B. The Role of Machine Learning in the Understanding and Design of Materials. Journal of the American Chemical Society **142**, 20273–20287, doi:10.1021/jacs.0c09105 (2020).
11. Baum, Z. J. *et al.* Artificial Intelligence in Chemistry: Current Trends and Future Directions. Journal of Chemical Information and Modeling **61**, 3197–3212, doi:10.1021/acs.jcim.1c00619 (2021).
12. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. Nature **559**, 547–555, doi:10.1038/s41586-018-0337-2 (2018).
13. Stephenson, N. *et al.* Survey of Machine Learning Techniques in Drug Discovery. Curr Drug Metab **20**, 185–193, doi:10.2174/1389200219666180820112457 (2019).
14. Khan, S. R., Al Rijjal, D., Piro, A. & Wheeler, M. B. Integration of AI and traditional medicine in drug discovery. Drug Discovery Today **26**, 982–992, doi:https://doi.org/10.1016/j.drudis.2021.01.008 (2021).
15. Rohall, S. L. *et al.* An Artificial Intelligence Approach to Proactively Inspire Drug Discovery with Recommendations. Journal of Medicinal Chemistry **63**, 8824–8834,

doi:10.1021/acs.jmedchem.9b02130 (2020).

16. Yi, Z. *et al.* Mapping Drug-Induced Neuropathy through In-Situ Motor Protein Tracking and Machine Learning. Journal of the American Chemical Society **143**, 14907–14915, doi:10.1021/jacs.1c07312 (2021).
17. Espinoza, G. Z., Angelo, R. M., Oliveira, P. R. & Honorio, K. M. Evaluating Deep Learning models for predicting ALK-5 inhibition. PLOS ONE **16**, e0246126, doi:10.1371/journal.pone.0246126 (2021).
18. Stokes, J. M. *et al.* A Deep Learning Approach to Antibiotic Discovery. Cell **180**, 688–702.e613, doi:10.1016/j.cell.2020.01.021 (2020).
19. Bess, A. *et al.* Artificial intelligence for the discovery of novel antimicrobial agents for emerging infectious diseases. Drug Discov Today, doi:10.1016/j.drudis.2021.10.022 (2021).
20. Kundu, S. *et al.* Enabling early detection of osteoarthritis from presymptomatic cartilage texture maps via transport-based learning. Proc Natl Acad Sci U S A **117**, 24709–24719, doi:10.1073/pnas.1917405117 (2020).
21. Tsigelny, I. F. Artificial intelligence in drug combination therapy. Brief Bioinform **20**, 1434–1448, doi:10.1093/bib/bby004 (2019).
22. Patel, L., Shukla, T., Huang, X., Ussery, D. W. & Wang, S. Machine Learning Methods in Drug Discovery. Molecules **25**, 5277, doi:10.3390/molecules25225277 (2020).
23. Gao, C. *et al.* Innovative Materials Science via Machine Learning. Advanced Functional Materials **32**, 2108044, doi:https://doi.org/10.1002/adfm.202108044 (2022).
24. Yin, Z.-W. *et al.* Advanced Electron Energy Loss Spectroscopy for Battery Studies. Advanced Functional Materials **32**, 2107190, doi:https://doi.org/10.1002/adfm.202107190 (2022).
25. Miljković, F., Rodríguez-Pérez, R. & Bajorath, J. Impact of Artificial Intelligence on Compound Discovery, Design, and Synthesis. ACS Omega, doi:10.1021/acsomega.1c05512 (2021).
26. Tkatchenko, A. Machine learning for chemical discovery. Nature Communications **11**, 4125, doi:10.1038/s41467-020-17844-8 (2020).
27. Safiri, S. *et al.* Global, regional and national burden of osteoarthritis 1990–2017: a systematic analysis of the Global Burden of Disease Study 2017. Annals of the Rheumatic Diseases **79**, 819–828, doi:10.1136/annrheumdis-2019-216515 (2020).
28. Buckwalter, J. A., Mankin, H. J. & Grodzinsky, A. J. Articular cartilage and osteoarthritis. Instructional course lectures **54**, 465–480 (2005).
29. Bajpayee, A. G., Wong, C. R., Bawendi, M. G., Frank, E. H. & Grodzinsky, A. J. Avidin as a model for charge driven transport into cartilage and drug delivery for treating early stage post-traumatic osteoarthritis. Biomaterials **35**, 538–549, doi:10.1016/j.biomaterials.2013.09.091 (2014).
30. Geiger, B., Grodzinsky, A. & Hammond, P.
31. Geiger, B. C., Wang, S., Padera, R. F., Grodzinsky, A. J. & Hammond, P. T. Cartilage-penetrating nanocarriers improve delivery and efficacy of growth factor treatment of osteoarthritis. Science Translational Medicine **10**, eaat8800, doi:10.1126/scitranslmed.aat8800 (2018).

32. J.W.G., J. & J.W.J., B. *Glucocorticoid therapy. In Kelley's Textbook of Rheumatology 7th edition*. 870–874 (Elsevier Saunders, 2005).
33. Perni, S. & Prokopovich, P. Poly-beta-amino-esters nano-vehicles based drug delivery system for cartilage. Nanomedicine **13**, 539–548, doi:10.1016/j.nano.2016.10.001 (2017).
34. Perni, S. & Prokopovich, P. Optimisation and feature selection of poly-beta-amino-ester as a drug delivery system for cartilage. J Mater Chem B **8**, 5096–5108, doi:10.1039/c9tb02778e (2020).
35. Green, J. J., Langer, R. & Anderson, D. G. A combinatorial polymer library approach yields insight into nonviral gene delivery. Acc Chem Res **41**, 749–759, doi:10.1021/ar7002336 (2008).
36. Burger, S. V. *Introduction to machine learning with R: rigorous mathematical analysis*. (2018).
37. Friedman, J., Hastie, J. & Tibshirani, R. *The elements of statistical learning*. (2009).
38. Russo, D. P., Zorn, K. M., Clark, A. M., Zhu, H. & Ekins, S. Comparing Multiple Machine Learning Algorithms and Metrics for Estrogen Receptor Binding Prediction. Mol Pharm **15**, 4361–4370, doi:10.1021/acs.molpharmaceut.8b00546 (2018).
39. Korotcov, A., Tkachenko, V., Russo, D. P. & Ekins, S. Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. Mol Pharm **14**, 4462–4475, doi:10.1021/acs.molpharmaceut.7b00578 (2017).
40. Fan, Y. *et al.* Investigation of Machine Intelligence in Compound Cell Activity Classification. Mol Pharm **16**, 4472–4484, doi:10.1021/acs.molpharmaceut.9b00558 (2019).
41. Guan, X. *et al.* Clinical and inflammatory features based machine learning model for fatal risk prediction of hospitalized COVID-19 patients: results from a retrospective cohort study. Ann Med **53**, 257–266, doi:10.1080/07853890.2020.1868564 (2021).
42. Kuhn, M. & Johnson, K. *Applied Predictive Modeling*. (2013).
43. Zhou, Z. H. *Ensemble Methods: Foundations and Algorithms*. (Chapman and Hall/CRC, 2012).
44. Fanourgakis, G. S., Gkagkas, K., Tylianakis, E. & Froudakis, G. E. A Universal Machine Learning Algorithm for Large-Scale Screening of Materials. Journal of the American Chemical Society **142**, 3814–3822, doi:10.1021/jacs.9b11084 (2020).
45. Chen, J. *et al.* Machine Learning Aids Classification and Discrimination of Noncanonical DNA Folding Motifs by an Arrayed Host:Guest Sensing System. Journal of the American Chemical Society **143**, 12791–12799, doi:10.1021/jacs.1c06031 (2021).
46. Jang, J., Gu, G. H., Noh, J., Kim, J. & Jung, Y. Structure-Based Synthesizability Prediction of Crystals Using Partially Supervised Learning. Journal of the American Chemical Society **142**, 18836–18843, doi:10.1021/jacs.0c07384 (2020).
47. Guo, Y. *et al.* Machine-Learning-Guided Discovery and Optimization of Additives in Preparing Cu Catalysts for CO2 Reduction. Journal of the American Chemical Society **143**, 5755–5762, doi:10.1021/jacs.1c00339 (2021).
48. Xie, Y. *et al.* Machine Learning Assisted Synthesis of Metal–Organic Nanocapsules. Journal of the American Chemical Society **142**, 1475–1481, doi:10.1021/jacs.9b11569 (2020).

49. Hatakeyama-Sato, K., Tezuka, T., Umeki, M. & Oyaizu, K. AI-Assisted Exploration of Superionic Glass-Type Li + Conductors with Aromatic Structures. Journal of the American Chemical Society **142**, 3301–3305, doi:10.1021/jacs.9b11442 (2020).
50. Tiihonen, A. *et al.* Predicting Antimicrobial Activity of Conjugated Oligoelectrolyte Molecules via Machine Learning. Journal of the American Chemical Society **143**, 18917–18931, doi:10.1021/jacs.1c05055 (2021).
51. Capasso Palmiero, U., Kaczmarek, J. C., Fenton, O. S. & Anderson, D. G. Poly(β-amino ester)-co-poly(caprolactone) Terpolymers as Nonviral Vectors for mRNA Delivery In Vitro and In Vivo. Adv Healthc Mater **7**, e1800249, doi:10.1002/adhm.201800249 (2018).
52. Moskowitz, J. S. *et al.* The effectiveness of the controlled release of gentamicin from polyelectrolyte multilayers in the treatment of Staphylococcus aureus infection in a rabbit bone model. Biomaterials **31**, 6019–6030, doi:10.1016/j.biomaterials.2010.04.011 (2010).
53. Anderson, D. G., Lynn, D. M. & Langer, R. Semi-Automated Synthesis and Screening of a Large Library of Degradable Cationic Polymers for Gene Delivery. Angewandte Chemie International Edition **42**, 3153–3158, doi:https://doi.org/10.1002/anie.200351244 (2003).
54. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.. (2019).
55. Kuhn, M. The caret Package. Journal of Statistical Software **28** (2012).

Figures

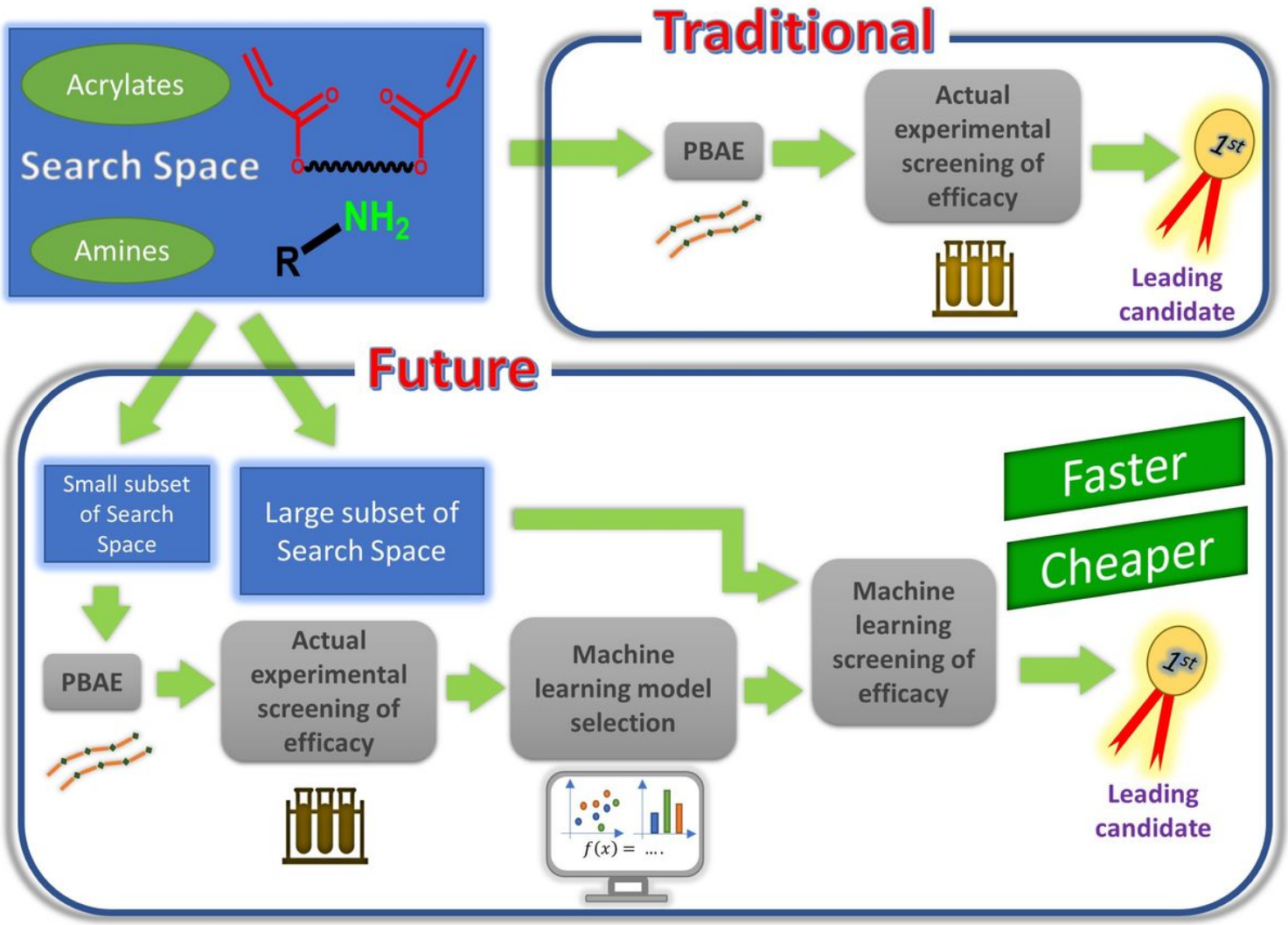

Figure 1

Schematic representation of machine learning driven drug development process.

Figure 2

Comparison of the different performance of the tested algorithms on the train (blue) and test set (red).

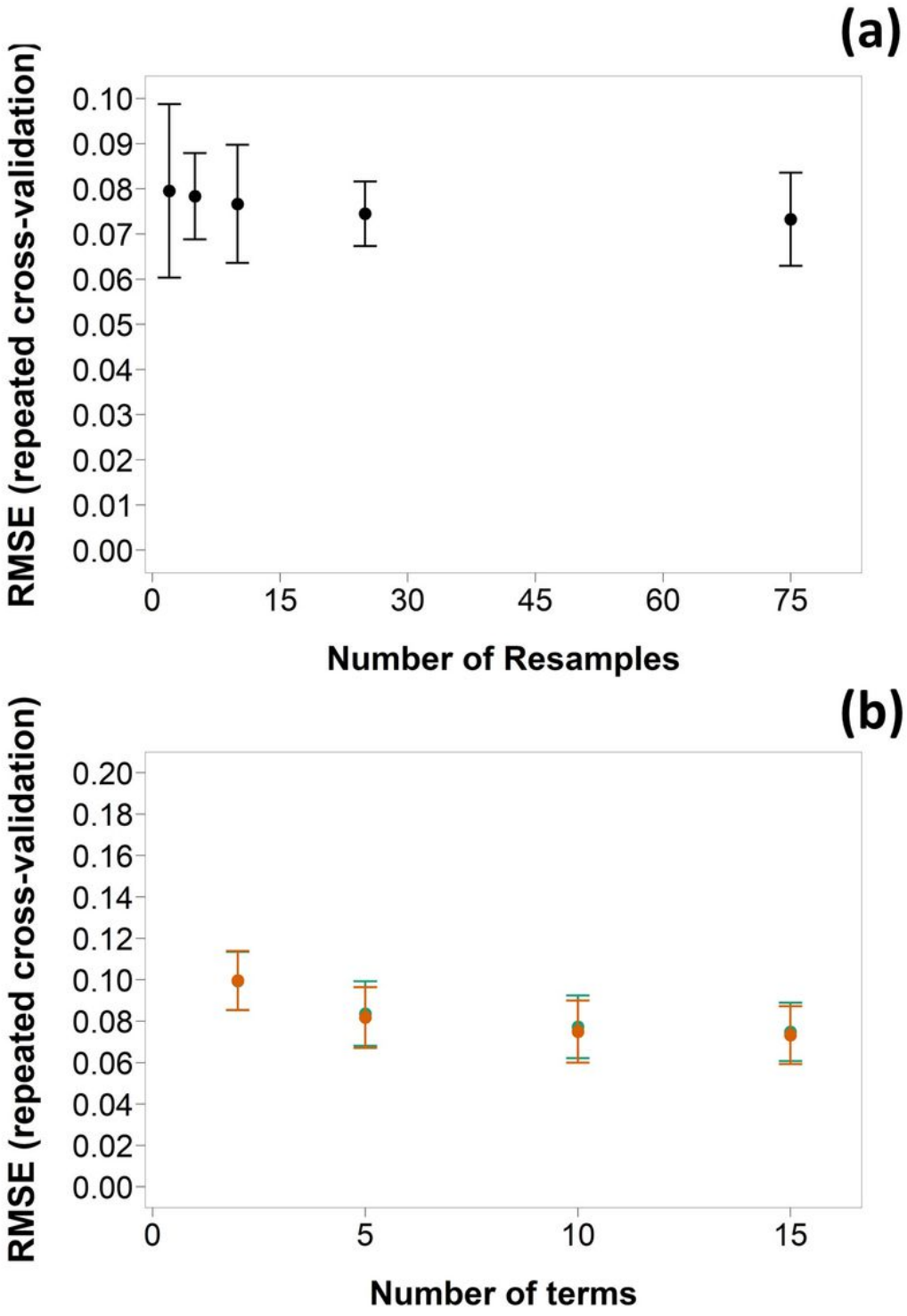

Figure 3

Relation for bag MARS models between RMSE (mean ± SD) for 10-fold cross validation repeated 3 times and (a) number of resamples and (b) number of terms and degree of correlation (n = 1 ■, n = 2 ■) (number of resamples =75).

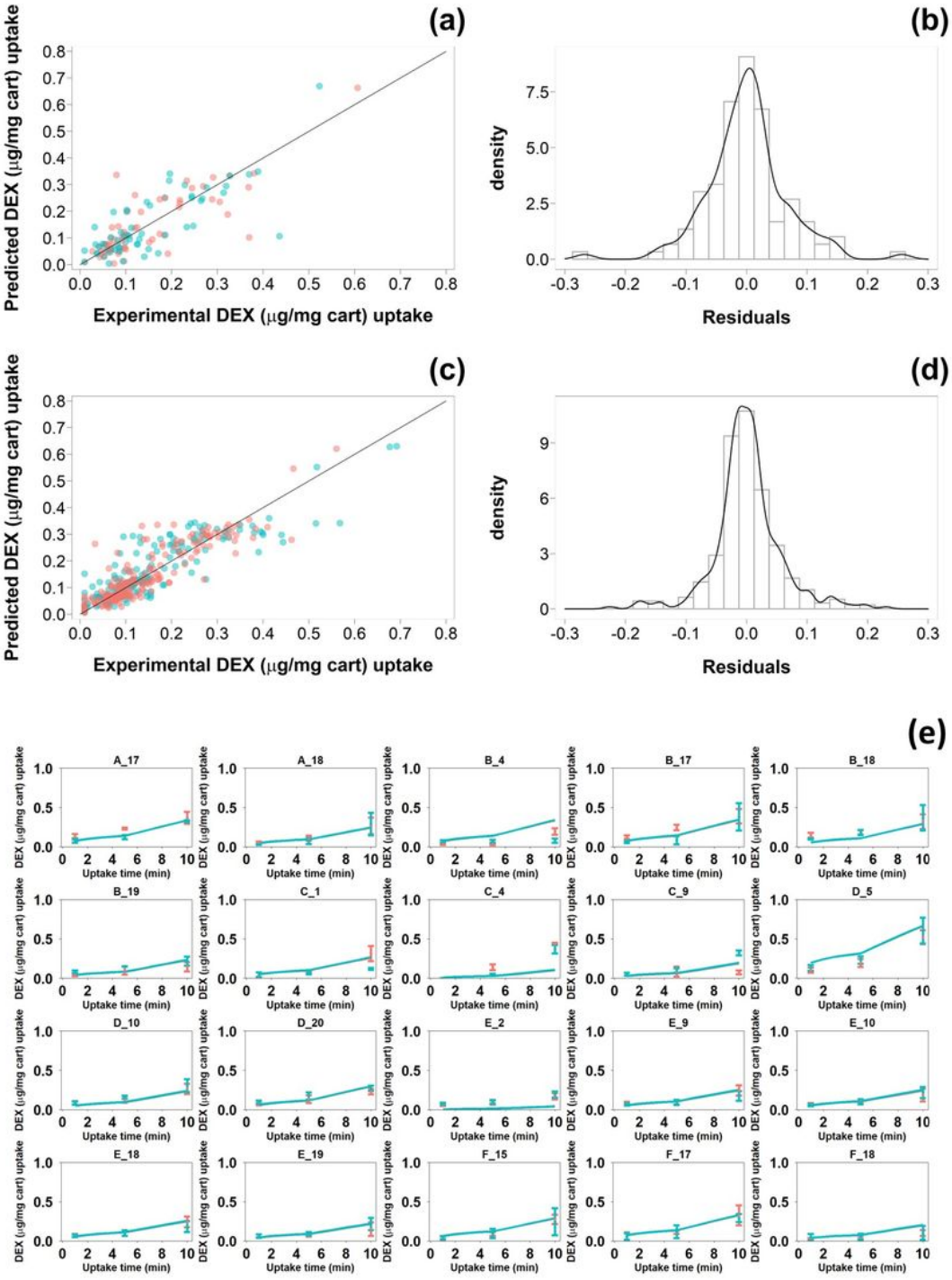

Figure 4

Comparision of predicted and experimental DEX uptake for PBAE-DEX (endcapped with e-1 ■ and e-2 ■) in the test (a) and train set (c); distributioin of residuals of DEX uptake predictions for PBAE-DEX in the test (b) and train set (d). Comparison of time dependent DEX uptake (mean ± SD) in cartilages predicted by optimised bag MARS model for PBAE in the test dataset (e).

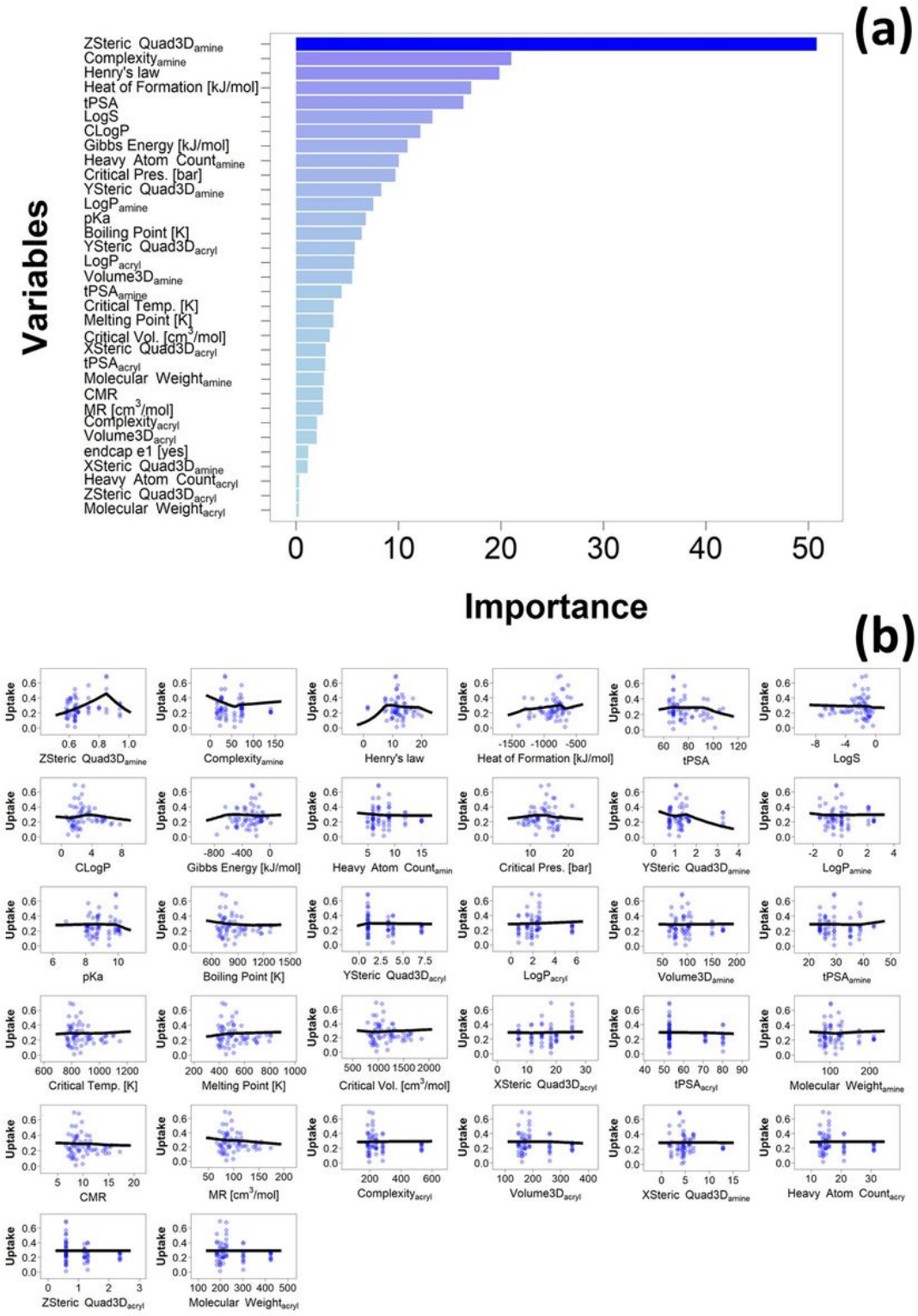

Figure 5

Variable importance in optimised bagged MARS model (a) and partial dependency plot of optimised bagged MARS model compared to experimentally obtained data for 10 min uptake of DEX into cartilage (b).

Figure 6

Heatmap of predicted ratio of DEX uptake for PBAE endcapped with e1 conjugated with DEX over commercial formulation of DEX after 10 min of exposure and structure of PBAE with predicted drug uptake superior to experimental found candidate.

Figure 7

Heatmap of predicted ratio of DEX uptake for PBAE during round2 endcapped with e1 conjugated with DEX over commercial formulation of DEX after 10 min of exposure and structure of PBAE with predicted drug uptake superior to best candidate in round1.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- suppinfo.docx