European Journal of Pharmaceutical Sciences 191 (2023) 106562

Contents lists available at ScienceDirect

European Journal of Pharmaceutical Sciences

journal homepage: www.elsevier.com/locate/ejps

Advancing algorithmic drug product development: Recommendations for machine learning approaches in drug formulation

Jack D. Murray [a], Justus J. Lange [a,b], Harriet Bennett-Lenane [a], René Holm [c], Martin Kuentz [d], Patrick J. O'Dwyer [a], Brendan T. Griffin [a,*]

[a] School of Pharmacy, University College Cork, Cork, Ireland
[b] Roche Pharmaceutical Research & Early Development, Pre-Clinical CMC, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, Grenzacherstrasse 124, Basel, Switzerland
[c] Department of Physics, Chemistry and Pharmacy, University of Southern Denmark, Campusvej 55, Odense 5230, Denmark
[d] School of Life Sciences, University of Applied Sciences and Arts Northwestern Switzerland, Muttenz CH 4132, Switzerland

ARTICLE INFO

Keywords:
Machine learning
Artificial intelligence
Computational pharmaceutics
Drug formulation
Data-driven modelling
Property prediction

ABSTRACT

Artificial intelligence is a rapidly expanding area of research, with the disruptive potential to transform traditional approaches in the pharmaceutical industry, from drug discovery and development to clinical practice. Machine learning, a subfield of artificial intelligence, has fundamentally transformed *in silico* modelling and has the capacity to streamline clinical translation. This paper reviews data-driven modelling methodologies with a focus on drug formulation development. Despite recent advances, there is limited modelling guidance specific to drug product development and a trend towards suboptimal modelling practices, resulting in models that may not give reliable predictions in practice. There is an overwhelming focus on benchtop experimental outcomes obtained for a specific modelling aim, leaving the capabilities of data scraping or the use of combined modelling approaches yet to be fully explored. Moreover, the preference for high accuracy can lead to a reliance on black box methods over interpretable models. This further limits the widespread adoption of machine learning as black boxes yield models that cannot be easily understood for the purposes of enhancing product performance. In this review, recommendations for conducting machine learning research for drug product development to ensure trustworthiness, transparency, and reliability of the models produced are presented. Finally, possible future directions on how research in this area might develop are discussed to aim for models that provide useful and robust guidance to formulators.

1. Introduction

Machine learning (ML) has never been more accessible to pharmaceutical scientists. Computational performance and high speed computing has increased in an exponential way since the 1970s (Leiserson et al., 2020). There are further modern algorithms available for ML and there is great interest in applying these to drug product development, with the number of published data-driven modelling studies growing year by year (Wang et al., 2021). In parallel, product development has become more challenging as drug discovery has also evolved technologically over the previous decades (Lou et al., 2023; Park et al., 2022). Rational small molecule drug design produces lead molecules based on a combination of screening, detailed structural knowledge of a biological target and the molecule, where the biological target is typically lipophilic (Bergström and Yazdanian, 2016; Doytchinova, 2022). High lipophilicity and increased molecular weight are reflected in molecules that emerge from combinatorial chemistry and high-throughput based drug discovery, explaining why drug candidates in the drug development pipeline exhibit poor biopharmaceutical properties. Formulation scientists are increasingly tasked with formulating poorly water soluble and poorly permeable drug candidates (Keserü and Makara, 2009; Vinarov et al., 2021). Early developability

Abbreviations: AdaBoost, adaptive boosting; AI, artificial intelligence; DOI, digital object identifier; FAIR, findable, accessible, interoperable, reusable; LASSO, least absolute shrinkage and selection operator; LIME, local interpretable model-agnostic explanations; ML, machine learning; MeSH, medical subject headings; PURL, persistent uniform resource locater; PMML, predictive model markup language.

* Corresponding author.
E-mail address: Brendan.Griffin@ucc.ie (B.T. Griffin).

https://doi.org/10.1016/j.ejps.2023.106562
Received 15 May 2023; Received in revised form 9 July 2023; Accepted 7 August 2023
Available online 9 August 2023

assessments can prompt scientists to abandon highly lipophilic lead candidates due to poor solubility or at least highlight the development risk, while other development hurdles can also limit clinical translation as high lipophilicity may also lead to suboptimal drug distribution, metabolism, and/or toxicology (Agarwal et al., 2022). Moreover, the trend towards an increasing molecular size often limits drug permeability, which is a particular issue for some promising new drug classes (Edmondson et al., 2019).

The needs of these difficult-to-formulate molecules are not optimally met by the conventional iterative trial-and-error approach to formulation design (Kuentz et al., 2016). A shift from empirical drug formulation development to computationally supported pharmaceutics may mitigate the unpredictability associated with the classical approach to drug product design. *In silico* models used in computational pharmaceutics can be broadly divided into theory-driven models including simulations that are based on physical sciences, and data-driven modelling based on empirical pattern recognition and algorithmic relationships (Kuentz and Bergström, 2021; Wang et al., 2021). Artificial Intelligence (AI) exists within data-driven modelling and is a term used to describe a computer system that simulates human intelligence. ML is a special case of AI where algorithms improve performance through exposure to data without being explicitly programmed (Bini, 2018; Poole et al., 1998).

At the fundamental level, ML tasks aim to construct models that generalise accurately to new, unseen data by utilising statistical inferences from previously observed input data. The effective construction of a model requires careful consideration of the balance between bias, representing the propensity to underfit the data, and variance, indicating the tendency to overfit the data. This trade-off between bias and variance is at the core to the development of accurate machine learning models and remains a critical concern throughout the model development process. With supervised learning, the most common form of ML seen in drug product development, a type of algorithm is presented with a set of inputs mapped to known outputs. The goal is to train the model to predict output values for previously unseen inputs. Supervised ML can be further subdivided into regression, to predict a continuous numerical outcome, and classification, to determine which class an unseen case belongs to, depending on the nature of the output (Bannigan et al., 2021). Conversely, the inputs to unsupervised learning algorithms are not mapped to an output. Instead, the algorithm clusters similar data together and finds empirical associations. Unsupervised learning can be also used as a preliminary step before the data are passed to a supervised learning algorithm (Maltarollo et al., 2015). The term semi-supervised learning is applied when aspects of both supervised and unsupervised learning are combined (Raschka and Kaufman, 2020). A final form is reinforcement learning, which is rarely used in drug product development, and entails continuous improvement as an algorithm rewards desired outcomes and penalises undesired outcomes (Elbadawi et al., 2021; Mak and Pichika, 2019).

Despite the increasing number of publications applying ML, the lack of availability of recommendations for scientists on how to fit data-driven models may result in a tendency towards suboptimal modelling practices. The resulting models often fall short of the expectation of robust predictions. For example, a model to predict the optimum formulation for a molecule with challenging biopharmaceutical properties that is overfit to its training data will not provide reliable predictions once deployed. Therefore, suboptimal computational pharmaceutics models may in the future cause more costly development failures than the benchtop experiments they aspire to replace. While research has not fully addressed suboptimal modelling as a potential issue for drug product development, this phenomenon has been extensively documented for small molecule drug discovery (Schaduangrat et al., 2020) and clinical medicine (Ellis et al., 2022).

Fig. 1 identifies sources of suboptimality in ML modelling in drug product development research, namely: irreproducible modelling; high-dimensional modelling; overlooking existing datasets; a lack of *in vivo* data/modelling; and a reliance on black boxes. Failure to publish source code and data makes it difficult for a third-party researcher to reproduce the ML model and leverage off previous research in the field. In terms of inputs, both the dimensionality and quantity of instances used to fit models may be unsuitable, in particular for modelling a small number of samples that each have many descriptors. Small datasets have dominated research to date, producing models that may be trained with a sufficient number of examples or have limited domains of applicability. The promise of ML to deliver robust predictions of quality formulation attributes, including *in vivo* attributes such as bioavailability, is currently still not fulfilled. Moreover, increased model interpretability is needed from a regulatory perspective in the framework of building quality into the final product (*i.e.*, Quality by Design) (Yu et al., 2014).

Fig. 1. Barriers to the increased adoption of ML for drug product development.

This review presents the current state-of-play of ML research in drug product development. While parallels are drawn with ML for drug discovery and clinical medicine, the focus here is on the path from drug substance to licensed drug product. The possible limitations of the current methodologies as outlined in Fig. 1 are presented in detail and discussed considering the current literature, and finally, strategies are suggested with proposed best practice for future research in this area.

2. Risks of irreproducible modelling

Reproducibility is central to the scientific method. *In silico* experiments are no more immune to the current 'crisis of reproducibility' than laboratory studies (Gibney, 2022). Reproducibility has no single agreed definition and in this review, Gundersen's definition of reproducibility for ML is applied: '*…the ability of independent investigators to draw the same conclusions from an experiment by following the documentation shared by the original investigators*' (Gundersen, 2021). It is important to note that reproducibility encompasses both data availability (*i.e.* the need for original data sharing for an independent investigator to replicate the model) and performance (*i.e.* the need for the model to have similar performance for new, real-world data as it showed on its test set).

From a mathematical viewpoint, there is inherent randomness with ML modelling. For example, the data being used in supervised learning to fit a model is often randomly split into test, train, and validation sets, and different models can be produced depending on how the data is partitioned. Similarly, in the case of neural networks, the initial synaptic weights are small random numbers (Agatonovic-Kustrin and Beresford,

2000). Notwithstanding such causes of randomness, a third-party researcher should be able to independently reproduce the same ML model given sufficient information in the form of data and code, especially in cases where the random seed used is specified. A robust framework for reproducibility is essential to maximise adoption of ML more widely for drug product development. However, in many cases ML models in the published pharmaceutics literature often do not detail sufficient information for a third party to replicate the results.

ML models cannot be readily reproduced if the data used to train them are not published in a machine readable format (Haymond and Master, 2022; Heil et al., 2021). Data-driven modelling is classically seen as a black-box process and publication of data is a key step that can be taken to enhance transparency. The FAIR (Findable, Accessible, Interoperable, and Reusable) Principles for Data Management and Stewardship should underpin the publication of all code, data, and metadata associated with predictive modelling for drug product development (Wilkinson et al., 2016). All data should, for example, have their own globally unique and persistent identifier, such as a digital object identifier (DOI) or persistent uniform resource locater (PURL) (Gundersen et al., 2018). Thus, data available only after contacting the author or publisher is not considered open in the strictest sense. Data published as tables within PDF documents are not interoperable as they are not readable by popular libraries for data analysis and manipulation, such as NumPy (Harris et al., 2020) and pandas (McKinney, 2010). Even if textual descriptors of the modelling methodology are published, these alone cannot be used directly to produce an identical model as small details are often lost without code (Gundersen and Kjensmo, 2018).

Clearly, data-driven research should have raw and processed data published as standard. To illustrate the lack of data availability in the drug product development applications, a simple Medical Subject Headings (MeSH) term search ('drug formulation' AND 'machine learning') was employed to identify a set of recent publications that can be analysed in detail. The aim was not to identify all relevant papers, but instead to produce a reasonably sized sample of studies that could be independently obtained by another researcher in a reproducible manner. By applying these MeSH terms in a search up to December 2022 (Supplementary Material 1), 27 relevant research articles were identified, of which 22 did not publish the data used to fit their models (Fig. 2).

There are some legitimate reasons not to publish all data explicitly, including cases where input data are proprietary. Drug product development models that are used in-house at pharmaceutical companies are likely to be trained with the firm's own investigational compounds and thus the training data is not appropriate for publication. This is reflected in the FAIR Principles, as there is no requirement for data to be open to be FAIR. It has further been suggested that studies in these instances can still share model predictions and data labels. The foundations of 'open as possible' and 'closed as necessary' must be balanced on a case-by-case basis in the context of pharmaceutical science datasets (Haibe-Kains et al., 2020).

It is also essential for code to be published alongside data. The traditional framework of scientific publishing, where the method used is described textually in a research article, does not best meet the needs of ML research (Schwab et al., 2000). It is highlighted across sources that describing the modelling in a 'Materials and Methods' section alone is generally insufficient for the output to be reproducible (Gundersen, 2021; Haibe-Kains et al., 2020). Uploading code to an open-source repository not only increases transparency but allows for other researchers to put this code into production without having to write it themselves. It has also been argued that even the publication of source code does not go far enough for deep learning algorithms – as changes in hardware that the code is executed on (*e.g.* thread count and random seed selection) were demonstrated in certain circumstances to eventually alter outcomes by an order of magnitude (Crane, 2018). This should be kept in mind with the aim to use end-to-end modelling along the full spectrum of drug discovery through to clinical trials, which means that initial model predictions of physicochemical parameters or formulation properties can lead to substantial error propagation if used as input for subsequent models that predict a product's *in vivo* performance. Code should also be annotated such that researchers without technical knowledge of the project can broadly interpret the modelling process. In

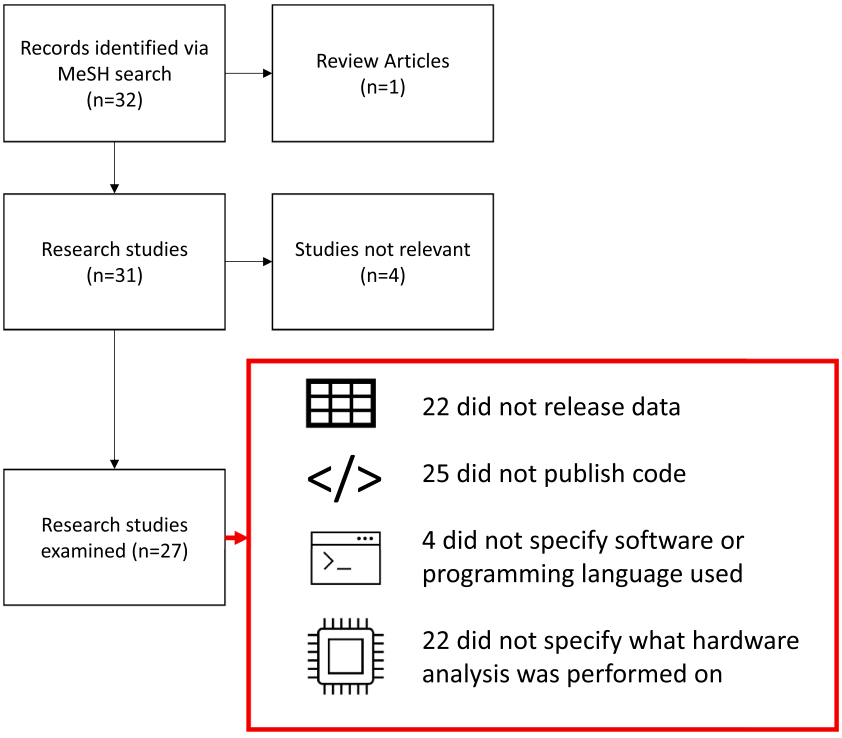


Fig. 2. A flow diagram summarising the results of the MeSH search strategy ('drug formulation' AND 'machine learning') used to identify a set of relevant papers for analysis.

the same set of 27 articles identified via MeSH searching, 25 did not include their code, and 22 did not specify the hardware that the analysis was performed on. Only one of the 27 articles specified all three of data, code, and hardware. The purpose of this analysis is to demonstrate the need for increased data and code sharing, rather than to criticise the reproducibility of individual papers. These studies still represent advancements in our understanding of what pharmaceutically relevant data is amenable to modelling. While the reliability of the models is more difficult to assess without supporting data, it is important to differentiate between reproducibility challenges due to inadequate documentation and those arising from the scientific method.

Further adding to the reproducibility crisis across disciplines is the phenomenon of data leakage. Just as clinical trials are blinded, so too should creation of a ML model be completely blind to the test data that will be used to assess performance. Data leakage occurs when information about the test set is unintentionally exchanged with the training set, causing a model to perform well on its test set but then not give accurate predictions when it is used for real-world prediction. Accuracy may be overestimated on the test set for a number of reasons, such as if the data in the test set was included when selecting predictor variables, or if the test set is not representative of real-world data, both of which are examples of data leakage (Kaufman et al., 2012). In a literature survey across 17 fields, Kapoor and Narayanan identified 329 papers affected by data leakage, and proposed a taxonomy of eight types of leakage to assist elimination of it (Kapoor and Narayanan, 2022).

In the absence of code and data, it is impossible to audit the full extent of data leakage in computational pharmaceutics research. Keeping with Kapoor and Narayanan's classification system, 'feature selection on training and test set' is an unintended source of potential leakage in recent articles. In one study which describes a prediction model for ternary cyclodextrin complexes, the model inputs (or 'features') are chosen using all of the data, not just the training data (Li et al., 2022). Therefore, information about the test set that the model will be assessed against was used to improve the model's performance. Leakage can be more subtle, however: in the prediction of tensile strength and disintegration time of tablets using Partial Least Squares analysis, where it was reported that data needed to be centred and scaled, but did not specify if this happened before or after the data was partitioned into training, testing, and validation sets (Hayashi et al., 2021). Variables should be rescaled using statistics calculated from the training set only, to prevent the leaking of information about the test set to the ML algorithm.

Kapoor and Narayanan advocate for the use of 'model info sheets' to accompany published predictive models. These prompt scientists to consider all possible sources of data leakage, and ensure training data is not inadvertently contaminated with information from the test set. A template for model info sheets is included with their arXiv preprint and includes questions such as process used to select the test set, and how the test-train split was maintained at each model evaluation step (Kapoor and Narayanan, 2022). Such a framework has clear potential to enhance reproducibility of ML models in the field of drug formulation. A deployed ML model is only expected to make reliable predictions within the space it was trained on. Describing the 'applicability domain' of a model can establish the boundaries of its predictive capabilities and assist in preventing predictions beyond this range, as such predictions would be unreliable. Considering the vast diversity in the chemical space encountered, it is advocated to report the applicability domain for ML models. In essence, these methods revolve around reporting the similarity of a new encountered sample compared to the predictor space encountered within the training set. There are several methods to quantify the applicability domain, including identifying the most important features and examining their distribution and extremes in the training set, as well as using Principal Component Analysis to compare the position of new samples with the training set in a lower-dimensional space (Netzeva et al., 2005).

The predictive models built by drug product development scientists cannot be used in practice for new data if the algorithms cannot be reproduced. In cases where models are not constructed using code but instead by statistical software, other means to facilitate the exchange of algorithms must be identified. Predictive Model Markup Language (PMML), an Extensible Markup Language that is both machine and human readable, can be used to facilitate interoperability and score new examples (Guazzelli et al., 2009). Another means of broadening the adoption of predictive models is the development of user-friendly applications. A successful example is M3DISEEN.com, where users can quickly generate *in silico* predictions relating to fused deposition modelling three dimensional printing of an oral tablet (Elbadawi et al., 2020). The model card framework proposed by Mitchell et al. can also be used as a tool by drug development scientists to disseminate key information about their model, particularly to a non-expert audience (Mitchell et al., 2019). The reader is also directed to recent computational reproducibility standards for life science researchers proposed by Heil et al. (2021).

To avoid discrepancies between how well the model performs on its test set and how well the model performs on real-world data, careful considerations during training of the model must be made to ensure its generalisation capability. The strategy employed to guarantee this will depend on the algorithm under evaluation and the structure of the data. Hyperparameters (*e.g.* the number of trees in a random forest or the regularization strength of a least absolute shrinkage and selection operator (LASSO) regression model) control the behaviour of an ML model and are specified by the user before training. They control the architecture and functionality of the model, and significantly impact model performance. The optimal hyperparameter values should be tuned by using cross-validation schemes on the training data to avoid train-test leakage, while at the same time considering different structural aspects of the data within the training set. This process facilitates training of the algorithm without sacrificing data and inducing train-test leakage. Cross-validation works by partitioning the training data into multiple subsets and iteratively training the model on different combinations. The final estimator evaluation can be conducted on the test set to ensure true model performance (Probst et al., 2019).

3. Risks of high-dimensional modelling

While the focus of this review is on data-driven computational modelling, mechanistic modelling, where equations that describe physical, chemical, and biological processes are solved or approximated by computational approaches, have also seen success in guiding drug formulation design. Molecular dynamics, computational fluid dynamics, and discrete element modelling to name but a few similarly achieve the goal of reducing the need for laboratory experiments and accelerating development (Mehta et al., 2019). These mechanistic models often have well-defined inputs as there are equations that underpin the phenomena under investigation. This is not equally clear in the case of data-driven modelling. It may be possible in some cases to select predictor variables based on theoretical equations (*e.g.* octanol-water partition coefficient and melting point in the case of solubility prediction) (Jain and Yalkowsky, 2001; Kuentz and Bergström, 2021). For many modelling problems, however, it is unclear which features are most predictive of the response variable. A key goal of ML is to reveal such previously unknown patterns or correlations in the data so this aspect of variable importance should be addressed in the modelling strategy.

Data is typically presented to ML algorithms as lists of values called vectors. Each value in the vector represents a different input variable, also known as a feature, that describes the data. Each case in the dataset is represented by its own vector, and the number of features (variables) in the vector is called its dimensionality. While it may seem counterintuitive at first, a vector with more features will not necessarily generate better predictions, as too many dimensions make it difficult for ML algorithms to produce models that generalise well. This is due to the 'Curse of Dimensionality', which refers to the loss of underlying patterns in the data as dimensions are added and the feature space becomes

increasingly sparse (Berisha et al., 2021).

This is especially relevant to drug product development datasets that can often have far more descriptors (*i.e.* features) than samples (*i.e.* vectors) (Jain and Zongker, 1997; Raudys and Jain, 1991). A classic example is drug solubility prediction, where thousands of molecular descriptors, each representing an additional dimension, are immediately available from the literature or via rapid calculation. However, the labour and cost associated with solubility studies limits the number of samples, and therefore training cases for ML (Bergström and Larsson, 2018; Kuentz and Bergström, 2021). Preliminary solubility estimation is crucial at the early stages of drug product development and as such it is in the interest of the field to use algorithms to streamline this process. Therefore, the benefits of modelling in lower dimensions and regularization techniques, including increased interpretability, lower computational cost of modelling, and, in some cases, increased accuracy by reducing noise and overfitting, should be explored (Jia et al., 2022).

For unsupervised clustering algorithms, such as *k*-means clustering, each vector is placed in a multi-dimensional space and the algorithm attempts to group similar cases together. Extra dimensions increase the average pairwise distance between data points and can result in loss of clear clustering seen in lower dimension problems (Altman and Krzywinski, 2018; Cunningham and Delany, 2022; Jain and Chandrasekaran, 1982). Similarly, for supervised learning algorithms like some artificial neural networks, the available information relative to the number of dimensions decreases with each additional independent variable. Rather than finding a general 'line of best fit', ML algorithms tend to overfit when there are many variables but only a few samples, and instead 'connect the dots.' The model will then fail to generalise on data that it did not encounter during training, resulting in unexpected and inaccurate predictions in practice (Schittenkopf et al., 1997).

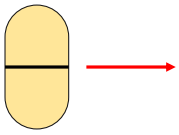
Categorical nominal inputs (for example, route of administration, excipients used during formulation, dosage form) are not readable by some ML algorithms in their native form, with the notable exception of tree models. As such, they must be transformed into a numerical format, with the most widespread technique being 'one-hot encoding.' One-hot encoded features are described as a list of binary values as described in Fig. 3. This introduces complexity to a modelling problem, as an extra dimension must be added for each category included in the dataset. Seemingly straightforward problems quickly acquire many dimensions, necessitating a large sample size to model effectively (Hastie et al., 2009).

ML is generally accepted to work best with large datasets, but with the expense associated with drug product development studies, a key factor to consider is the number of training cases that are required. Topliss and Costello observed that when there were fewer than five cases for each independent variable in a simple linear regression model, any correlations observed were likely due to chance (Topliss and Costello, 1972). This would suggest a linear growth in the number of samples needed for each additional dimension, but this is not always the case in practice. For example, each additional one-hot encoded variable results in exponential growth of required sample size, as for n binary variables, there are 2^n possible combinations of categories (see Fig. 3). Furthermore, introducing many binary dimensions to a small dataset can result in category combinations appearing in testing/validation sets that the model did not encounter in training. Models trained on limited data show strong bias as the model overfits and performance 'improves' due to noise instead of signal (Hastie et al., 2009).

One essential step in ML modelling that is often overlooked is feature engineering – using data science and domain expertise to manipulate raw data into more rational algorithm inputs (Bengio et al., 2013). The exception to this is certain deep learning methods where the model is fit using the raw data and feature extraction occurs autonomously. This autonomous approach has seen applications in chemical-chemical interaction potential (Kwon and Yoon, 2017) and drug design (Monteiro et al., 2021), amongst others. Dimensionality reduction without loss of predictive accuracy, however, remains a key objective for most models.

It is computationally cheaper and faster to train models with fewer dimensions. Additionally, models with fewer inputs can be advantageous as some features may be laborious to determine or cumbersome to calculate. Dimensionality reduction has been shown in some instances to increase accuracy by reducing noise (Jia et al., 2022). There is no general-purpose approach to optimising features. The calculation of derived features and grouping of similar categorical inputs can greatly mitigate the 'Curse of Dimensionality'. Principal component analysis and partial least squares regression in particular have seen extensive applications to reduce dimensionality in drug product development datasets (Ferreira and Tobyn, 2015).

Another obvious approach is to exclude inputs which do not make a meaningful contribution to output (Steppe and Bauer, 1997). In order to filter redundant features, the strength of the univariate relationship between each descriptor and the target output can be quantified by various importance metrics, such as mutual information or the Pearson coefficient. Importance analysis must only be performed with training data to prevent leakage. Similarly, 'greedy engineering methods' add or reduce the number of variables used by a model in a stepwise manner



	Propylene Glycol	Alpha Tocopherol	Lecithin	Methylparaben	Long-Chain Triglycerides	Polysorbate 80
Propylene Glycol	1	0	0	0	0	0
Alpha Tocopherol	0	1	0	0	0	0
Lecithin	0	0	1	0	0	0
Methylparaben	0	0	0	1	0	0
Long-Chain Triglycerides	0	0	0	0	1	0
Polysorbate 80	0	0	0	0	0	1

Fig. 3. *Simple illustration of one-hot encoding.* The figure above shows a fictitious lipid-based formulation. One-hot encoding creates a binary dummy variable for each category. Every excipient is given its own dummy variable. As the number of excipients increases, so too does the number of dummy variables and, therefore, dimensionality. If there are too many dummy variables, the feature space becomes sparse with many zeroes, and synergistic combinations of excipients may not be modelled effectively. One-hot encoding can be used in many other situations relevant to drug product development, such as to represent SMILES strings, atom types, and routes of administration.

based on importance and have seen success in optimising models for drug product development (Bennett-Lenane et al., 2021). While these methodologies reduce dimensionality, they ignore cases where inputs may have synergistic/antagonistic effects (Chandrashekar and Sahin, 2014; Guyon and Elisseeff, 2003). In the context of drug product development, however, domain expertise can offer insight into which features are naturally meaningful and which are expected to be redundant. Another approach to feature selection and dimension reduction is regularization such as L1 and L2 regularization applied by LASSO and Ridge Regression respectively. By adding a penalty term to the cost function that these models are trying to optimise, sparse coefficient values for each feature can be obtained, avoiding challenges inherent to overfitting and limited interpretability (Demir-Kavuk et al., 2011).

To illustrate the merits of feature engineering, consider a study which describes the use of a random forest model to predict the extent of agglomeration in pharmaceutical particulate systems using nine input features (Sinha et al., 2021). An importance analysis after the model was constructed and revealed that five of the nine features had a negligible effect on mean squared error of the model. By applying a feature selection approach and constructing the model again using only the four most important predictors, accuracy may improve (Sinha et al., 2021). Comparably, many of the 21 inputs in a study to predict tensile strength and disintegration time of tablets did not contribute meaningfully to the model's output (Hayashi et al., 2021). This study also produced two fully interpretable decision trees using only the top seven most important inputs from their two earlier models. The benefits of this methodology in terms of dimensionality reduction are obvious. The added advantage of interpretability gain is discussed in detail under '6. *Reliance on Black Boxes'*.

Determining the best inputs or mathematical transformations of these inputs is computationally challenging. It has been shown mathematically that the only method to determine the optimum subset of descriptors is an exhaustive search of each possible subset, but in high dimensional drug product development datasets this may not be feasible (Cover and Van Campenhout, 1977). A modelling problem with 20 features, for example, would require 1,048,575 solutions to be constructed (2^n possible sets minus the null set). This also ignores the potential for dimensionality reduction or derived features. The features that minimise the error for one algorithm may not be the same for another, or even the same algorithm with a different internal architecture. Moreover, feature selection algorithms can be unstable for small datasets that have many dimensions, meaning that each iteration of the algorithm can give different results (Dernoncourt et al., 2014). Even though the solutions set provided by feature selection algorithms may be suboptimal, the final model can still be of high predictive quality.

4. Possible opportunities with existing data

The 'Curse of Dimensionality' emphasises the need for sufficient high-quality data to effectively model a given problem. Computational chemists have assembled immense datasets from which data-driven modelling can be performed. In one library, for example, the docking of 170 million make-on-demand compounds against two biological targets was used to identify novel highly potent agents (Lyu et al., 2019). Schneider et al. propose that ML approaches could use big data like this to identify promising ways in which libraries can be expanded further into the most relevant drug-like chemical space (Schneider et al., 2020). In the clinical sciences, the availability of labelled big data has fuelled applications of ML to diagnostic image processing and ECG/EEG interpretation (Topol, 2019). The potential of pharmaceutical consortia to gather experimental datasets and synthesise structured data from the existing body of unstructured digital information has yet to be fully realised.

It has been demonstrated already in this review that small sample size modelling with many independent variables leads to a sparse feature space and makes overfitting likely. Despite this, the prevailing approach to ML in the drug product development literature involves conducting small-scale experiments for the purpose of modelling the results obtained. An alternative approach is to exploit insights from the existing regulatory or literature data where available. A current issue to this end is that there are limited databases or otherwise machine-readable data that are relevant to the formulation scientist for such ML efforts. However, with increasing numbers of open data publications and efforts in pharmaceutical consortia to gather larger datasets, this current situation is about to change. International campaigns of data sharing and generation, including the Simcyp consortium (Jamei et al., 2013), OrBiTo (Lennernäs et al., 2014), PEARRL (Kuentz et al., 2021) and most recently InPharma (Reppas et al., 2023), will continue to result in databases that can be used for ML research. This would form a basis for a shift in focus from production to consumption of data in ML, which has the potential to uncover underlying patterns in current knowledge and prevent wasteful duplication of experiments.

It is promising that the rate of data generation generally is increasing year on year, with some of this data likely to be meaningful for drug product development. International Data Corporation project that the datasphere, the total digital data in the world, will amount to 175 zettabytes by 2025 (Reinsel et al., 2018). The rate of data generation far exceeds our current extraction of useful information from it, with most data being unstructured and challenging to analyse. This is echoed by regulators. The Heads of Medicines Agencies-European Medicines Agency Big Data Taskforce define big data as 'extremely large datasets which may be complex, multi-dimensional, unstructured and heterogeneous…' (HMA-EMA Big Data Taskforce, 2019). Data mining, the process of extracting information from big data using ML and statistics, is needed for pattern and anomaly detection. The Food and Drug Administration routinely uses data mining tools to process large safety report datasets (Duggirala et al., 2016), and are expanding this to other aspects of the drug lifecycle as part of the Knowledge-aided Assessment and Structured Application System (Yu et al., 2019). Literature examples, however, of data mining being applied to data for drug product development are limited.

Whether a study produces or consumes data depends mainly on the availability of machine-readable data. For example, there is potential for data to be scraped from the regulatory datasphere in ML modelling campaigns to predict the bioavailability of marketed formulations (Bennett-Lenane et al., 2022). It is important to recognise that the data of interest to drug product development scientists may not be as easily accessible in the same way as small molecule libraries due to a lack of high throughput experimental methods. Smaller datasets will likely be easier to convert into a machine-readable format, in contrast to big data where significant cleaning may be required to appropriately format the information. Smaller datasets that are produced from combined efforts in, for example, a consortium would have fewer confounding parameters from non-standardized methods or other noise factors that are commonly included in gathered unstructured datasets.

None of the studies uncovered in this review exploited big data, which is again likely due to the limited availability of big data relevant to drug formulation. Many studies, however, demonstrated that literature data can be effectively mined. Li et al. combined data from multiple papers to train ML models to optimise the formulation of drug/cyclodextrin/polymer complexes (Li et al., 2022). Wang et al. highlighted the challenges of combining data from different sources in this manner, such as heterogeneity of structure (Wang et al., 2021). Often with this approach, the available data differs between studies, requiring fields with missing values to be either dropped or have these numbers imputed (Emmanuel et al., 2021). The strategy for handling missing data in this case was guided by a combination of statistics and domain knowledge. The prediction of subcutaneous monoclonal antibody bioavailability has also been guided exclusively from data consumption. Lou and Hageman constructed a database of monoclonal antibody bioavailability with 47 potential predictors that were obtained from the literature (Lou and Hageman, 2021). Some studies combined experimental and literature

data, *e.g.* a model to predict tablet properties from material properties used a dataset that contained elements of data consumption (*e.g.* by expanding on existing material libraries via literature search) and production (*e.g.* determination of tensile strength and disintegration time experimentally) (Hayashi et al., 2021).

Data fuels ML, and drug product development lacks the enormous, structured data libraries available for other scientific disciplines such as drug design and medical image processing. Searches of open dataset collections yield far more results for medical images and drug design than drug formulation. The 'drug-like' chemical space of molecules that conform to Lipinski's Rule of Five for oral absorption (Lipinski et al., 2001) has been estimated to be in excess of 10^{60} (Bohacek et al., 1996; Reymond, 2015). This is especially relevant in the field of excipient design. Defining a chemical space for excipient design is not as clear given the structural diversity and the many roles they fulfil in the various dosage forms. This is complicated further by excipients that cannot be defined as a single chemical structure, such as many polymers and lipid-based formulation excipients (Rowe et al., 2009). Existing explorations of excipient chemical space have been isolated to specific classes of agents, including freeze-drying excipients (Meng-Lund et al., 2019) and stabilising molecules for antibodies (Tosstorff et al., 2020), and not aided by ML.

The ability of computers to generate this chemical data far exceeds the human capacity to perform the same task. Investigations of the extent of the small molecule chemical universe to date have been predominantly automated. The GDB-17 database enumerates 166.4 billion molecules of up to 17 atoms of carbon, nitrogen, oxygen, sulphur, and halogens (Reymond, 2015). At the time of the publication of GDB-17, PubChem, the largest publicly available collection of chemical information, listed only 2.5 million compounds that conform to the same restrictions as the GDB-17 database (Ruddigkeit et al., 2012). Machines could ultimately outperform humans in their power to aggregate existing observations in the drug product development datasphere. A vision of formulation sciences would be to scrape unstructured textual data from the drug product development datasphere/literature with a subsequent transformation into structured data for ML. Large volumes of human-readable semi-structured text from regulatory documents, safety reports, etc., can be systematically separated into segments, with useful segments being retained and added to a database (Aho, 1990). Natural language processing applied to drug development goes a step further than regular expression searching. Rather than being rules based, natural language processing is an AI technique for the analysis of human language (Nadkarni et al., 2011). A recent review by Bhatnagar et al. illustrated the applications, challenges, and future opportunities of natural language processing in drug development (Bhatnagar et al., 2022). Machine vision problems (Ficzere et al., 2022; Thite et al., 2022) may also benefit from large automatically generated image databases from which training examples may be obtained.

5. Lack of *in vivo* modelling

Identifying *in vitro* and *in silico* models for *in vivo* prediction is a fundamental objective of pharmaceutical science. Although *in silico* modelling has traditionally focused on the theory-driven approach, such as Physiologically Based Pharmacokinetic Modelling, AI and the surge in data has sparked expanding interest in data-driven representations of physical and biological processes. To date, the applications of ML to drug product development have concentrated on optimising factors such as processability, product stability, etc., rather than enhancing *in vivo* performance. From the set of articles identified by MeSH searching, only one was trained using *in vivo* data to predict an *in vivo* outcome. Lou and Hageman sourced subcutaneous bioavailability values of monoclonal antibodies from regulatory documents and literature to derive a relationship between product properties and bioavailability (Lou and Hageman, 2021). This permits efficient estimation of subcutaneous bioavailability for many monoclonal antibodies and formulation strategies without the need for benchtop data. Rich libraries that integrate published data, molecular descriptors, and *in vivo* human outcomes, such as the dataset of intravenous pharmacokinetic parameters reported by Lombardo et al. (2018), provide a promising framework for predictive model design.

There is limited feasibility to employ a data production pipeline for ML studies predicting *in vivo* outputs for given formulations. Human studies are expensive and require ethical approval. Animal studies are also used to gain *in vivo* data for drug product development. However, there are ethical and cost considerations associated with collection of animal data for modelling. In some cases, such as for bioavailability determination, results do not necessarily correlate across species, complicating the construction of a ML model (Musther et al., 2014). Additionally, outside the personalised medicine domain, traditional formulation decisions are made at the population level to identify a one-size-fits-all dosage form, which creates a scarcity of patient level data at a specific subpopulation level *e.g.* special population level (Trenfield et al., 2018).

There is also significant variability in how clinical trial results are reported (Comets and Zohar, 2009), with patient-level data often omitted. Thus, a ML model fit using population statistics will learn differences between formulations for a whole population instead of learning from the variability one might obtain by training a model with individual data points. Consequently, Lou and Hageman's (2021) model, for example, could not be used for patient-level predictions of exposure. *In vivo* data tend to show high inter-occasion and inter-individual variability, in comparison to laboratory or manufacturing data that are less noisy. Investigational drug molecules identified via screening processes that favour poorly water soluble molecules in particular are likely to show variable oral absorption and hepatic metabolism (Di et al., 2012). This presents unique obstacles for intelligent algorithms as individual prediction, like many biological processes (Sejdić and Lipsitz, 2013), has a stochastic nature.

Given the considerable variability of *in vivo* performance, it is crucial that the populations from which pharmacokinetic parameters are obtained are representative of the patients that will use the formulation. It is well documented that ML can amplify bias, with many prominent examples in clinical medicine of models that do not perform well on underrepresented patient subpopulations (Mehrabi et al., 2022). When collecting values such as bioavailability, food effect, and inter-individual variability from the datasphere, any bias in the design of the clinical trial to determine these figures will be integrated into the model. The complexity deepens when different trials have different levels of representation, variable sample sizes, or when this information is not reported. Algorithmic approaches to lessen this bias, including oversampling, undersampling, and adversarial debiasing, are of limited benefit here because frequently the data from underrepresented groups is lacking completely (Vokinger et al., 2021). In trials where there are small numbers of underrepresented populations and parameters are reported as trial averages only, data relating to these individuals cannot be oversampled. A model card (Mitchell et al., 2019) accompanying ML models of *in vivo* outcomes should fully describe the data used to train the algorithm, biases associated with that data, and the population for which the prediction is expected to be valid.

Another source of bias stemming from the nature of studies in regulatory documents is the lack of data on failed drug products. Formulations that already fail to provide sufficient oral bioavailability in preclinical animal testing are typically not tested in early clinical trials for a comparison with a bio-enabling formulation. Trial failures of drug candidates abandoned for commercial reasons are often lost from the publicly accessible drug product development datasphere. Recent analysis shows that the compliance with reporting requirements on ClinicalTrials.gov, the world's largest clinical trials registry, is poor (DeVito et al., 2020). Even for investigational products that fail at late-stage clinical trials, a majority of studies are not published in peer-reviewed journals (Hwang et al., 2016).

Furthermore, the marketed product may not be the optimum one in terms of biopharmaceutical performance. Fenofibrate (Ling et al., 2013) and cyclosporin (Mueller et al., 1994), for example, both saw improved formulations with enhanced bioavailability and more reliable pharmacokinetics brought to market following approval of the parent formulation. Formulation decisions are made for ease of manufacturing, commercial reasons, and patient preference as well as *in vivo* concerns, and often the result is a balance of these sometimes competing factors. The formulation that maximises exposure, for example, may not be scalable to manufacture, (Stegemann et al., 2007) or there may be a company tradition or strategy to favour an alternative bio-enabling approach (Kuentz et al., 2016). Such multi-objective optimisation problems typically do not have a solution that satisfies all objectives simultaneously. Rather, there are a range of possible solutions, where optimising for one outcome can result in a trade off with another (Narayanan et al., 2021). Even at the stage of drug discovery, multi-objective optimisation may aim to balance receptor affinity, aqueous solubility, metabolic stability, and other key pharmacokinetic and pharmacodynamic parameters (Schneider et al., 2020). A model fit using data from marketed drug products is therefore not trained with formulations that are optimum, but with formulations that are satisfactory to see clinical translation.

Despite efforts to construct biologically relevant algorithms, there are key outcomes for which there are insufficient *in vivo* data for modelling. *In vitro* surrogates of *in vivo* outcomes, where laboratory outcomes are used in place of clinical trials, can save costs and eliminate the ethical burden of human trials. There are instances in which *in vitro* studies have particular advantages compared to *in vivo* studies with a limited number of subjects. For example, *in vitro* studies can more suitably investigate bioequivalence for immediate release solid oral dosage forms, excluding the high risk of Type II error of *in vivo* studies (Polli, 2008).

Therefore, the question arises, should intelligent *in silico* tools based on *in vitro* data be considered valid for the prediction of *in vivo* outputs? On one hand, correlation is not strictly transitive and therefore an *in silico-in vitro* model does not necessarily imply an *in silico-in vivo* relationship (Sotos et al., 2009). A model that performs well at predicting drug solubility in biorelevant media is not guaranteed to represent *in vivo* solubility. *In vitro* methods with a high degree of biorelevance may still have limited predictability as the mechanistic nuances of the *in vivo* context are lost, but this is case-dependant (Vinarov et al., 2021). This is an intense area of research and the biorelevance and predictive capabilities of *in vitro* methods are constantly evolving. Data-driven, theory-driven, and *in vitro* methods can also be used in parallel. A study by Parrott et al. demonstrated that certain pharmacokinetic parameters of lipophilic drugs can be accurately predicted by integrating ML-predicted properties and *in vitro* data with Physiologically Based Pharmacokinetic models (Parrott et al., 2022). This is essentially an approach that combined data-driven and mechanistic modelling, which is a promising approach to the outlined challenges in predicting *in vivo* performance of drug products.

6. Reliance on black boxes

Not all ML models are equally transparent for human understanding and while theory-driven models with explicit equations should be fully transparent, a diametral type of model is often called a 'black-box'. The reliance on black box modelling further limits the widespread adoption of ML models for drug product development. The model takes inputs and returns outputs, but the internal workings of the algorithm are not readily interpretable. Deep learning is a key illustration of this: many hidden layers of neurons perform nonlinear transformations on unstructured data to produce abstract representations which are not seen by the human building the model (LeCun et al., 2015). While there are types of data, such as image and video processing, that require this deep learning approach (Chen et al., 2021), a majority of drug product development data has structured inputs for which an impact on output would ideally be modelled with interpretable ML. Interpretable ML is an approach where the relationship between input and output data can be clearly understood (Fig. 4). Interpretability is a spectrum and is case and domain specific, but generally accepted examples of interpretable algorithms include decision trees, scoring systems, Naïve Bayes classifiers, rule lists, and simple linear or logistic regression models (Du et al., 2019; Molnar et al., 2020).

A common misconception is that black boxes always outperform interpretable models. While again problem-specific, both interpretable and black box ML methods tend to show similar accuracy for tabular data where there is appropriate feature engineering. Rudin (2019) argues that when performing ML on a dataset, there are a number of different modelling approaches which will all have similar accuracy,

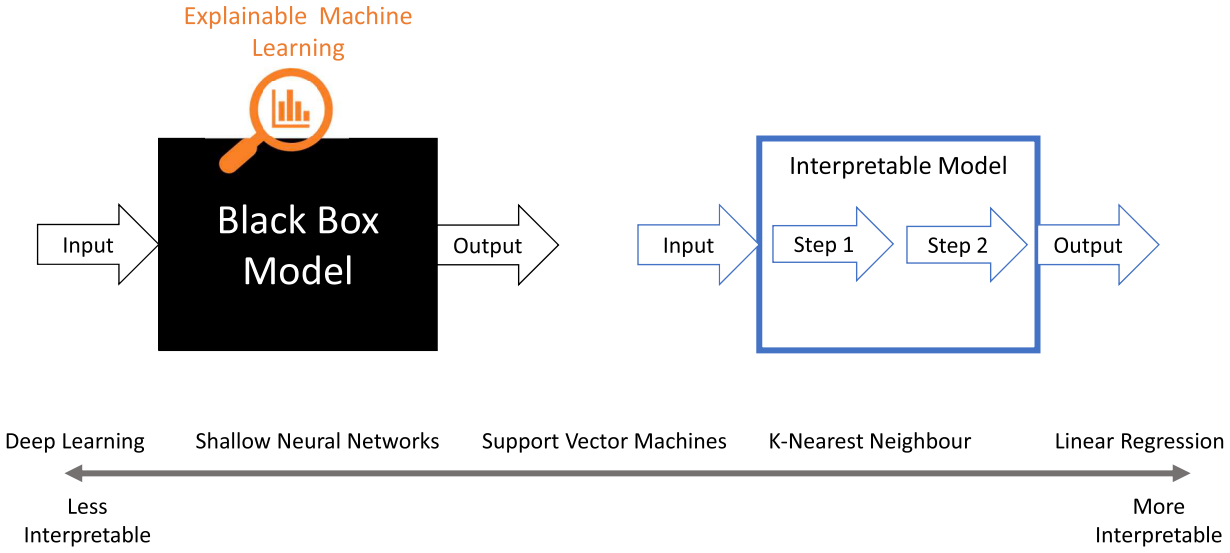


Fig. 4. A selection of commonly used ML algorithms arranged from least interpretable to most interpretable. The distinction between interpretable ML and explainable ML is also illustrated. Interpretable models, such as multiple linear regression, clearly show the relationship between input and output without the need for a separate explanation model. Black box models, such as deep neural networks, require explainable ML techniques to understand the relationship between input and output.

called the Rashomon set. This set of possible models is typically large when data is structured and inputs are naturally meaningful, as they are for most drug product development datasets. Rudin indicates in this set of equally predictive models, at least one is interpretable (Rudin, 2019).

The major advantage of interpretable ML models is that the effect of each input on the output under study is obvious, and thus the model is straightforward to both deploy and understand (Rudin, 2019). Although more human-readable and less computationally expensive to use in practice, these models are not necessarily easier to construct. Significant restrictions are placed on what the final model can look like, and therefore the production of an interpretable algorithm is a constrained optimisation problem, as opposed to unconstrained optimisation associated with black box models (Rudin et al., 2022). Rudin indicates also that much of the current ML talent is concentrated in deep learning, and that ML packages favour uninterpretable models. In short, accuracy may be easy to achieve but an optimum, sparse, interpretable algorithm that maintains this accuracy usually exists although it can be difficult to identify (Rudin, 2019). It was proven almost 50 years ago that it is computationally intractable to construct an optimal binary decision tree, as the problem is NP-Complete (nondeterministic polynomial time) (Hyafil and Rivest, 1976). In spite of this proof, the value of sparse and interpretable solutions is demonstrated by the volume of research into enhanced decision tree and rule list optimisation over the last 60 years (Angelino et al., 2018; Kotsiantis, 2013).

To understand the need for interpretable models, it is important to clarify the shortcomings of using explainable ML to understand black boxes. Unlike interpretable models, which are human readable, explainable ML aims to explain black box input-output relationships and justify how the model arrived at its prediction. Ranking of input importance is a simple example of explainable ML. The vast number of importance metrics, the inability to detect synergistic or antagonistic interactions between inputs, and difficulty in determining the direction of the relationship show how complicated black box explanations need to be. For further discussion the interested reader is directed to Rudin's perspective on the subject (Rudin, 2019).

In the context of formulation, predictive models are often built with the intention of inference rather than pure prediction. For example, to determine the best physicochemical properties of an excipient to solubilise a model drug, a study may be carried out where 100 solubilising excipients are examined. With an interpretable model, such as a decision tree, one can easily see what a theoretical 'best excipient' might look like. If explainable ML, as described above, were used in conjunction with a black box for the same problem, extensive trial and error with different input combinations may be required to optimise the excipient properties (Alarie et al., 2021).

Of the papers identified in this review there was clearly a higher prevalence of black box modelling, with only isolated examples of interpretable models. In a study of capping occurrence, both interpretable models (a decision tree and a logistic regression model) and black boxes (neural network and random forest) were constructed (Paul et al., 2021). All four models showed broadly similar accuracy for their test/validation sets, ranging from 90.7% for logistic regression to 96.4% for the neural network. The random forest, which contained an ensemble of 'up to 500 trees', performed only marginally better than the single interpretable decision tree on the same test set and reduced the number of misclassifications from 7/86 to 5/86. The best neural network misclassified 3/84 cases. Interestingly, the models disagreed over the direction of some of the relationships. A high interparticulate bonding strength was predicted to increase rate of capping for the logistic regression model, decrease rate of capping for the decision tree, and not significantly influence neural network prediction.

Lipton introduces the concept of decomposability to interpretability: *'each input, parameter, and calculation admits an intuitive explanation.'* (Lipton, 2017). Lou and Hageman's models for subcutaneous antibody bioavailability prediction include a decision tree, a random forest, and an Adaptive Boosting (AdaBoost) algorithm, amongst others (Lou and Hageman, 2021). Like the random forest, AdaBoost is an ensemble method that typically combines many weak learners, such as decision stumps, where subsequent stumps focus on incorrectly classified cases. All three algorithms showed the same performance on the validation set (separate to the training set and the test set used to tune hyperparameters), with 78% accuracy (Fig. 5). The decision tree outperformed the multilayer perceptron neural network, which correctly classified just 67% of validation cases. Despite the human readability of the decision tree architecture, the inputs generated are principal components of the original feature list in their drug product library. Although a major use case of Principal Component Analysis is to make high dimensional data more interpretable, in this case the strength and direction of relationship between an individual feature and model output is difficult to interpret. A decision tree split on molecular weight and intrinsic solubility score, for example, would be more interpretable than a decision tree split on linear combinations of all 47 inputs.

Examples of interpretability thus far have focused on classification problems, but interpretability is equally desirable for regression analysis. Interpretable regression models (a multiple linear regression model and a polynomial linear regression model) were fit to predict Young's modulus of pharmaceutical compacts. In both cases, there was an acceptable level of prediction ($R^2 = 0.81$ and $R^2 = 0.89$, respectively) and the effect of each independent variable on output is clear (Thomas et al., 2021). Even with explainable AI approaches like Shapley Additive Explanations (Lundberg and Lee, 2017) that justify how the black box arrives at individual predictions, the exact mathematical relationship between input and output cannot be fully understood. The Local Interpretable Model-Agnostic Explanations (LIME) algorithm (Ribeiro et al., 2016) is another explainable AI technique that creates a locally faithful interpretable approximation for any black box, but again cannot be used

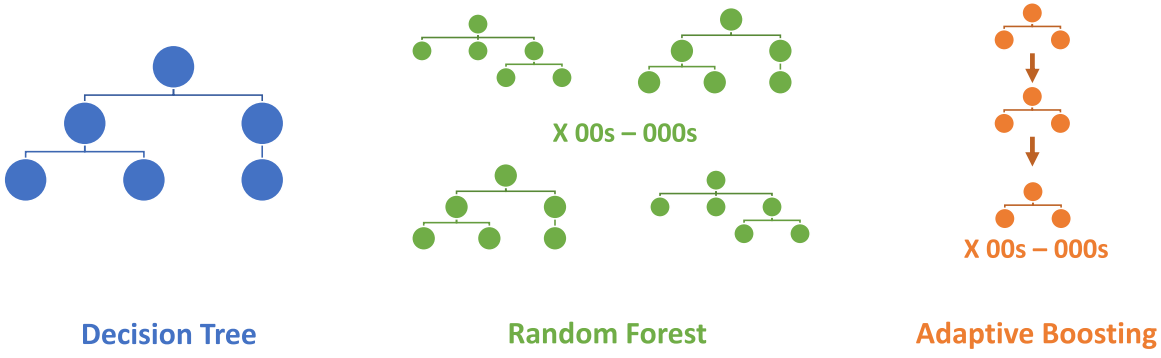


Fig. 5. Adapted from models produced by Lou and Hageman to predict subcutaneous bioavailability of an antibody formulation. It can be easily observed that only the decision tree's architecture is fully interpretable. All three models had equal accuracy (78%) on the same validation set, despite random forest and AdaBoost being ensemble methods with hundreds and sometimes thousands of decision trees contributing towards the final prediction.

to find a globally optimum formulation (assuming one exists).

It is impossible to generalise that all drug product development datasets are amenable to interpretable modelling. The need for deep learning methodologies is likely to grow as modelling problems involving non-tabular data such as images, natural language, and molecular representations become more prevalent. The surge in the use of graph neural network methods for molecular property prediction is one such example (David et al., 2020; Fang et al., 2022). Graph convolutional neural networks have shown excellent promise in the prediction of solubility, lipophilicity, and membrane affinity (Lee et al., 2022; Montanari et al., 2019). Even in the case of tabular data, however, Butler et al. offer an alternative hypothesis that ML may uncover scientific laws so complicated that they are beyond human understanding and thus require a black box to model (Butler et al., 2018). A thorough exploration of interpretable models may eliminate the need for a black box, and therefore models from across the spectrum of complexity (Fig. 4) should still be explored to identify the approach which fits the data best. Even if the final model is interpretable, previously unknown correlations and patterns in the data may be identified by more complex models. Further, by considering models of differing complexities, researchers can be more confident that the best model has been identified. Just as interpretability is a continuum, so too is explainability, with shallow models, for example, being more explainable than deep neural networks. In these cases, it is reasonable for the researcher to choose the model that minimises dimensionality and model complexity while preserving accuracy. Even for computer vision problems that generally require a convolutional neural network approach (Chen et al., 2021), efforts have been undertaken to make the algorithms more interpretable (Chen et al., 2019).

There is growing recognition in other fields, including clinical medicine, that interpretability is as critical as accuracy for important decisions. As previously demonstrated, intelligent algorithms can exaggerate bias present in the training set. Bias has been repeatedly reported in healthcare algorithms (Vokinger et al., 2021). By presenting the exact input-output relationship, bias introduced during the modelling step is easier to evaluate. The issue of confidence in predictive models for drug development was raised with respect to reproducibility, and the same reasoning holds true here. There is likely to be increased trust in an interpretable relationship than a black box, resulting in more widespread adoption of ML models for drug product development (Petch et al., 2022; Watson et al., 2019).

7. Recommendations

Suboptimal modelling practices have been documented in practically every field that ML has been applied to. This review demonstrates that drug product development is not an exception. In particular, the five broad areas of irreproducible modelling, high dimensionality, neglect of the datasphere, the *in vivo* gap, and the prevalence of black boxes were identified. As applied ML matures and drug product development becomes increasingly automated, these concerns should be remedied to improve the reliability of AI predictions. Confidence in data-driven models is necessary as ML removes the need for human input from progressively more critical and sophisticated tasks.

In Table 1, recommendations about how current modelling practices may be changed to advance the field of computational pharmaceutics considering the five limitations identified in this review are proposed by the authors. This guidance has been informed by the current state-of-art of both ML and drug product development science as described in this review. Under each heading, two readily implementable recommendations are made.

The recommendations in Table 1 can be broadly split into methodological recommendations (1.2, 2.2, 5.1, 5.2) and data-focused recommendations (1.1, 2.1, 3.1, 3.2, 4.1, 4.2). This is reflective of the fact that ML itself is driven by data. The way in which data is created and published is a critical determinant of the direction ML-informed drug product development will take. The recommendations proposed, including searching for interpretable solutions and data sharing, are straightforward to adopt and have the potential to significantly improve the quality of models produced. Moving towards optimal modelling practices from both a procedural and data management perspective has promising implications for the way in which medicines are made from molecules.

Table 1
Key recommendations for ML-informed drug product development.

Current Practice	Recommendation
1. Irreproducible modelling	1.1. Studies should publish data, annotated code, and specify hardware in a manner that conforms to the FAIR Principles. 1.2. Studies should release models with documentation, such as a model card and model info sheet, which describes the intended use of the model; data used during training and testing; and steps taken to identify data leakage.
2. High-dimensional modelling	2.1. Studies should assess the volume of data available before modelling and obtain further data where necessary. 2.2. Studies should perform appropriate feature engineering, including feature selection and dimensionality reduction.
3. Overlooking existing data	3.1. Studies should consider the availability of existing data before conducting benchtop experiments for generating data for ML to prevent duplication of experiments. 3.2. The potential for consortia to synthesise structured data from the existing body of unstructured digital information and, where no machine-readable data exists, gather large experimental datasets, should be investigated.
4. Lack of *in vivo* modelling	4.1. Studies should analyse the bias associated with the data used to fit ML models to predict *in vivo* outcomes and report this in the model's documentation. 4.2. Consortia should work towards creating large and open-source drug product libraries that integrate drug products with *in vivo* performance to drive ML in this area.
5. Preference for black boxes	5.1. Studies should thoroughly search for an interpretable model that performs as well as a black box model for problems involving structured drug product development data. 5.2. In cases where no interpretable model can be identified, studies should publish the results from all models explored and apply suitable explainable AI methods to the model selected.

8. Conclusion

The field of computational pharmaceutics has advanced significantly over the last decade, and for the full potential of ML informed drug product development to be realised, further improvements to the current approaches to data-driven modelling are desirable. In this review, recommendations to advance the field by improving the transparency, reliability, and accuracy of ML models were suggested. It was highlighted that suboptimal ML methodology is only one side of the story, with further progress to be made in terms of data stewardship in drug product development science. The expertise of the drug formulation scientist, along with results from *in vivo* and *in vitro* experiments, can be consolidated with machine-generated predictions to increasingly eliminate the guesswork at all stages of drug product development. Ultimately, just as computers now aid the discovery of molecules from receptors, an end-to-end *in silico* method for the generation of medicines from diseases based on ML may soon be realised.

Funding

JDM is funded by the Irish Research Council Government of Ireland Postgraduate Scholarship Programme grant number GOIPG/2022/

1580.

PJOD, MK, JJL, RH, and BTG are part of the InPharma European Training network, which has received funding under the Marie Skłodowska-Curie grant agreement No 955756.

CRediT authorship contribution statement

Jack D. Murray: Conceptualization, Writing – original draft. **Justus J. Lange:** Writing – review & editing. **Harriet Bennett-Lenane:** Supervision, Writing – review & editing. **René Holm:** Supervision, Writing – review & editing. **Martin Kuentz:** Supervision, Writing – review & editing. **Patrick J. O'Dwyer:** Supervision, Writing – review & editing. **Brendan T. Griffin:** Supervision, Writing – review & editing.

Data availability

No data was used for the research described in the article.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejps.2023.106562.

References

Agarwal, P., Huckle, J., Newman, J., Reid, D.L., 2022. Trends in small molecule drug properties: a developability molecule assessment perspective. Drug Discov. Today 27, 103366. https://doi.org/10.1016/j.drudis.2022.103366.

Agatonovic-Kustrin, S., Beresford, R., 2000. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. J. Pharm. Biomed. Anal. 22, 717–727. https://doi.org/10.1016/S0731-7085(99)00272-1.

Aho, A.V., 1990. Algorithms for finding patterns in strings. Algorithms and Complexity. Elsevier, pp. 255–300. https://doi.org/10.1016/B978-0-444-88071-0.50010-2.

Alarie, S., Audet, C., Gheribi, A.E., Kokkolaras, M., Le Digabel, S., 2021. Two decades of blackbox optimization applications. EURO J. Comput. Optim. 9, 100011 https://doi.org/10.1016/j.ejco.2021.100011.

Altman, N., Krzywinski, M., 2018. The curse(s) of dimensionality. Nat. Methods 15, 399–400. https://doi.org/10.1038/s41592-018-0019-x.

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., Rudin, C., 2018. Learning certifiably optimal rule lists for categorical data. J. Mach. Learn. Res. 18, 1–78.

Bannigan, P., Aldeghi, M., Bao, Z., Häse, F., Aspuru-Guzik, A., Allen, C., 2021. Machine learning directed drug formulation development. Adv. Drug Deliv. Rev. 175, 113806 https://doi.org/10.1016/j.addr.2021.05.016.

Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1798–1828.

Bennett-Lenane, H., Griffin, B.T., O'Shea, J.P., 2022. Machine learning methods for prediction of food effects on bioavailability: a comparison of support vector machines and artificial neural networks. Eur. J. Pharm. Sci. 168, 106018 https://doi.org/10.1016/j.ejps.2021.106018.

Bennett-Lenane, H., O'Shea, J.P., Murray, J.D., Ilie, A.R., Holm, R., Kuentz, M., Griffin, B. T., 2021. Artificial neural networks to predict the apparent degree of supersaturation in supersaturated lipid-based formulations: a pilot study. Pharmaceutics 13, 1398. https://doi.org/10.3390/pharmaceutics13091398.

Bergström, C.A.S., Larsson, P., 2018. Computational prediction of drug solubility in water-based systems: qualitative and quantitative approaches used in the current drug discovery and development setting. Int. J. Pharm. 540, 185–193. https://doi.org/10.1016/j.ijpharm.2018.01.044.

Bergström, C.A.S., Yazdanian, M., 2016. Lipophilicity in drug development: too much or not enough? AAPS J. 18, 1095–1100. https://doi.org/10.1208/s12248-016-9947-5.

Berisha, V., Krantsevich, C., Hahn, P.R., Hahn, S., Dasarathy, G., Turaga, P., Liss, J., 2021. Digital medicine and the curse of dimensionality. Npj Digit. Med. 4, 153. https://doi.org/10.1038/s41746-021-00521-5.

Bhatnagar, R., Sardar, S., Beheshti, M., Podichetty, J.T., 2022. How can natural language processing help model informed drug development?: A review. JAMIA Open 5, ooac043. https://doi.org/10.1093/jamiaopen/ooac043.

Bini, S.A., 2018. Artificial intelligence, machine learning, deep learning, and cognitive computing: what do these terms mean and how will they impact health care? J. Arthroplasty 33, 2358–2361. https://doi.org/10.1016/j.arth.2018.02.067.

Bohacek, R.S., McMartin, C., Guida, W.C., 1996. The art and practice of structure-based drug design: a molecular modeling perspective. Med. Res. Rev. 16, 3–50. https://doi.org/10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6.

Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O., Walsh, A., 2018. Machine learning for molecular and materials science. Nature 559, 547–555. https://doi.org/10.1038/s41586-018-0337-2.

Chandrashekar, G., Sahin, F., 2014. A survey on feature selection methods. Comput. Electr. Eng. 40, 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K., 2019. This looks like that: deep learning for interpretable image recognition. In: Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F.d', Fox, E., Garnett, R. (Eds.), Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), pp. 8930–8941.

Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S., Miao, Y., 2021. Review of image classification algorithms based on convolutional neural networks. Remote Sens. 13, 4712. https://doi.org/10.3390/rs13224712.

Comets, E., Zohar, S., 2009. A survey of the way pharmacokinetics are reported in published phase I clinical trials, with an emphasis on oncology. Clin. Pharmacokinet. 48, 387–395. https://doi.org/10.2165/00003088-200948060-00004.

Cover, T.M., Van Campenhout, J.M., 1977. On the possible orderings in the measurement selection problem. IEEE Trans. Syst. Man Cybern. 7, 657–661. https://doi.org/10.1109/TSMC.1977.4309803.

Crane, M., 2018. Questionable answers in question answering research: reproducibility and variability of published results. Trans. Assoc. Comput. Linguist. 6, 241–252. https://doi.org/10.1162/tacl_a_00018.

Cunningham, P., Delany, S.J., 2022. k-nearest neighbour classifiers - a tutorial. ACM Comput. Surv. 54, 1–25. https://doi.org/10.1145/3459665.

David, L., Thakkar, A., Mercado, R., Engkvist, O., 2020. Molecular representations in AI-driven drug discovery: a review and practical guide. J. Cheminform. 12, 56. https://doi.org/10.1186/s13321-020-00460-5.

Demir-Kavuk, O., Kamada, M., Akutsu, T., Knapp, E.W., 2011. Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features. BMC Bioinform. 12, 412. https://doi.org/10.1186/1471-2105-12-412.

Dernoncourt, D., Hanczar, B., Zucker, J.D., 2014. Analysis of feature selection stability on high dimension and small sample data. Comput. Stat. Data Anal. 71, 681–693. https://doi.org/10.1016/j.csda.2013.07.012.

DeVito, N.J., Bacon, S., Goldacre, B., 2020. Compliance with legal requirement to report clinical trial results on ClinicalTrials.gov: a cohort study. Lancet 395, 361–369. https://doi.org/10.1016/S0140-6736(19)33220-9.

Di, L., Fish, P.V., Mano, T., 2012. Bridging solubility between drug discovery and development. Drug Discov. Today 17, 486–495. https://doi.org/10.1016/j.drudis.2011.11.007.

Doytchinova, I., 2022. Drug design—past, present, future. Molecules 27, 1496. https://doi.org/10.3390/molecules27051496.

Du, M., Liu, N., Hu, X., 2019. Techniques for interpretable machine learning. Commun. ACM 63, 68–77. https://doi.org/10.1145/3359786.

Duggirala, H.J., Tonning, J.M., Smith, E., Bright, R.A., Baker, J.D., Ball, R., Bell, C., Bright-Ponte, S.J., Botsis, T., Bouri, K., Boyer, M., Burkhart, K., Steven Condrey, G., Chen, J.J., Chirtel, S., Filice, R.W., Francis, H., Jiang, H., Levine, J., Martin, D., Oladipo, T., O'Neill, R., Palmer, L.A.M., Paredes, A., Rochester, G., Sholtes, D., Szarfman, A., Wong, H.L., Xu, Z., Kass-Hout, T., 2016. Use of data mining at the food and drug administration. J. Am. Med. Inform. Assoc. 23, 428–434. https://doi.org/10.1093/jamia/ocv063.

Edmondson, S.D., Yang, B., Fallan, C., 2019. Proteolysis targeting chimeras (PROTACs) in 'beyond rule-of-five' chemical space: recent progress and future challenges. Bioorg. Med. Chem. Lett. 29, 1555–1564. https://doi.org/10.1016/j.bmcl.2019.04.030.

Elbadawi, M., Gaisford, S., Basit, A.W., 2021. Advanced machine-learning techniques in drug discovery. Drug Discov. Today 26, 769–777. https://doi.org/10.1016/j.drudis.2020.12.003.

Elbadawi, M., Muñiz Castro, B., Gavins, F.K.H., Ong, J.J., Gaisford, S., Pérez, G., Basit, A. W., Cabalar, P., Goyanes, A., 2020. M3DISEEN: a novel machine learning approach for predicting the 3D printability of medicines. Int. J. Pharm. 590, 119837 https://doi.org/10.1016/j.ijpharm.2020.119837.

Ellis, R.J., Sander, R.M., Limon, A., 2022. Twelve key challenges in medical machine learning and solutions. Intell.-Based Med. 6, 100068 https://doi.org/10.1016/j.ibmed.2022.100068.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., Tabona, O., 2021. A survey on missing data in machine learning. J. Big Data 8, 140. https://doi.org/10.1186/s40537-021-00516-9.

Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., Wang, F., Wu, H., Wang, H., 2022. Geometry-enhanced molecular representation learning for property prediction. Nat. Mach. Intell. 4, 127–134. https://doi.org/10.1038/s42256-021-00438-4.

Ferreira, A.P., Tobyn, M., 2015. Multivariate analysis in the pharmaceutical industry: enabling process understanding and improvement in the PAT and QbD era. Pharm. Dev. Technol. 20, 513–527. https://doi.org/10.3109/10837450.2014.898656.

Ficzere, M., Mészáros, L.A., Kállai-Szabó, N., Kovács, A., Antal, I., Nagy, Z.K., Galata, D. L., 2022. Real-time coating thickness measurement and defect recognition of film coated tablets with machine vision and deep learning. Int. J. Pharm. 623, 121957 https://doi.org/10.1016/j.ijpharm.2022.121957.

Gibney, E., 2022. Could machine learning fuel a reproducibility crisis in science? Nature 608, 250–251. https://doi.org/10.1038/d41586-022-02035-w.

Guazzelli, A., Zeller, M., Lin, W.C., Williams, G., others, 2009. PMML: an open standard for sharing models. R. J. 1, 60.

Gundersen, O.E., 2021. The fundamental principles of reproducibility. Philos. Trans. R. Soc. Math. Phys. Eng. Sci. 379, 20200210 https://doi.org/10.1098/rsta.2020.0210.

Gundersen, O.E., Gil, Y., Aha, D.W., 2018. On reproducible AI: towards reproducible research, open science, and digital scholarship in AI publications. AI Mag. 39, 56–68. https://doi.org/10.1609/aimag.v39i3.2816.

Gundersen, O.E., Kjensmo, S., 2018. State of the art: reproducibility in artificial intelligence. In: Proceedings of the AAAI Conference on Artificial Intelligence, p. 32. https://doi.org/10.1609/aaai.v32i1.11503.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182. https://doi.org/10.5555/944919.944968.

Haibe-Kains, B., Adam, G.A., Hosny, A., Khodakarami, F., , Massive Analysis Quality Control (MAQC) Society Board of Directors, Shraddha, T., Kusko, R., Sansone, S.A., Tong, W., Wolfinger, R.D., Mason, C.E., Jones, W., Dopazo, J., Furlanello, C.,

Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., Greene, C.S., Broderick, T., Hoffman, M.M., Leek, J.T., Korthauer, K., Huber, W., Brazma, A., Pineau, J., Tibshirani, R., Hastie, T., Ioannidis, J.P.A., Quackenbush, J., Aerts, H.J. W.L., 2020. Transparency and reproducibility in artificial intelligence. Nature 586, E14–E16. https://doi.org/10.1038/s41586-020-2766-y.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2.

Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Hayashi, Y., Nakano, Y., Marumo, Y., Kumada, S., Okada, K., Onuki, Y., 2021. Application of machine learning to a material library for modeling of relationships between material properties and tablet properties. Int. J. Pharm. 609, 121158 https://doi.org/10.1016/j.ijpharm.2021.121158.

Haymond, S., Master, S.R., 2022. How can we ensure reproducibility and clinical translation of machine learning applications in laboratory medicine? Clin. Chem. 68, 392–395. https://doi.org/10.1093/clinchem/hvab272.

Heil, B.J., Hoffman, M.M., Markowetz, F., Lee, S.I., Greene, C.S., Hicks, S.C., 2021. Reproducibility standards for machine learning in the life sciences. Nat. Methods 18, 1132–1135. https://doi.org/10.1038/s41592-021-01256-7.

HMA-EMA Big Data Taskforce, 2019. Phase II Report:'Evolving Data-Driven Regulation' (No. EMA/584203/2019). European Medicines Agency.

Hwang, T.J., Carpenter, D., Lauffenburger, J.C., Wang, B., Franklin, J.M., Kesselheim, A. S., 2016. Failure of investigational drugs in late-stage clinical development and publication of trial results. JAMA Intern. Med. 176, 1826. https://doi.org/10.1001/jamainternmed.2016.6008.

Hyafil, L., Rivest, R.L., 1976. Constructing optimal binary decision trees is NP-complete. Inf. Process. Lett. 5, 15–17. https://doi.org/10.1016/0020-0190(76)90095-8.

Jain, A., Zongker, D., 1997. Feature selection: evaluation, application, and small sample performance. IEEE Trans. Pattern Anal. Mach. Intell. 19, 153–158. https://doi.org/10.1109/34.574797.

Jain, A.K., Chandrasekaran, B., 1982. Dimensionality and sample size considerations in pattern recognition practice. Handbook of Statistics. Elsevier, pp. 835–855. https://doi.org/10.1016/S0169-7161(82)02042-2.

Jain, N., Yalkowsky, S.H., 2001. Estimation of the aqueous solubility I: application to organic nonelectrolytes. J. Pharm. Sci. 90, 234–252. https://doi.org/10.1002/1520-6017(200102)90:2<234::aid-jps14>3.0.co;2-v.

Jamei, M., Marciniak, S., Edwards, D., Wragg, K., Feng, K., Barnett, A., Rostami-Hodjegan, A., 2013. The simcyp population based simulator: architecture, implementation, and quality assurance. Silico Pharmacol. 1, 9. https://doi.org/10.1186/2193-9616-1-9.

Jia, W., Sun, M., Lian, J., Hou, S., 2022. Feature dimensionality reduction: a review. Complex Intell. Syst. 8, 2663–2693. https://doi.org/10.1007/s40747-021-00637-x.

Kapoor, S., Narayanan, A., 2022. Leakage and the reproducibility crisis in ML-based science.

Kaufman, S., Rosset, S., Perlich, C., Stitelman, O., 2012. Leakage in data mining: formulation, detection, and avoidance. ACM Trans. Knowl. Discov. Data 6, 1–21. https://doi.org/10.1145/2382577.2382579.

Keserü, G.M., Makara, G.M., 2009. The influence of lead discovery strategies on the properties of drug candidates. Nat. Rev. Drug Discov. 8, 203–212. https://doi.org/10.1038/nrd2796.

Kotsiantis, S.B., 2013. Decision trees: a recent overview. Artif. Intell. Rev. 39, 261–283. https://doi.org/10.1007/s10462-011-9272-4.

Kuentz, M., Bergström, C.A.S., 2021. Synergistic computational modeling approaches as team players in the game of solubility predictions. J. Pharm. Sci. 110, 22–34. https://doi.org/10.1016/j.xphs.2020.10.068.

Kuentz, M., Holm, R., Elder, D.P., 2016. Methodology of oral formulation selection in the pharmaceutical industry. Eur. J. Pharm. Sci. 87, 136–163. https://doi.org/10.1016/j.ejps.2015.12.008.

Kuentz, M., Holm, R., Kronseder, C., Saal, C., Griffin, B.T., 2021. Rational selection of bio-enabling oral drug formulations – a PEARRL commentary. J. Pharm. Sci. 110, 1921–1930. https://doi.org/10.1016/j.xphs.2021.02.004.

Kwon, S., Yoon, S., 2017. DeepCCI: end-to-end deep learning for chemical-chemical interaction prediction. In: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics. Presented at the BCB '17: 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. Boston Massachusetts USA. ACM, pp. 203–212. https://doi.org/10.1145/3107411.3107451.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539.

Lee, S., Lee, M., Gyak, K.W., Kim, S.D., Kim, M.J., Min, K., 2022. Novel solubility prediction models: molecular fingerprints and physicochemical features *vs* graph convolutional neural networks. ACS Omega 7, 12268–12277. https://doi.org/10.1021/acsomega.2c00697.

Leiserson, C.E., Thompson, N.C., Emer, J.S., Kuszmaul, B.C., Lampson, B.W., Sanchez, D., Schardl, T.B., 2020. There's plenty of room at the top: what will drive computer performance after Moore's law? Science 368, eaam9744. https://doi.org/10.1126/science.aam9744.

Lennernäs, H., Aarons, L., Augustijns, P., Beato, S., Bolger, M., Box, K., Brewster, M., Butler, J., Dressman, J., Holm, R., Julia Frank, K., Kendall, R., Langguth, P., Sydor, J., Lindahl, A., McAllister, M., Muenster, U., Müllertz, A., Ojala, K., Pepin, X., Reppas, C., Rostami-Hodjegan, A., Verwei, M., Weitschies, W., Wilson, C., Karlsson, C., Abrahamsson, B., 2014. Oral biopharmaceutics tools – time for a new initiative – an introduction to the IMI project OrBiTo. Eur. J. Pharm. Sci. 57, 292–299. https://doi.org/10.1016/j.ejps.2013.10.012.

Li, J., Gao, H., Ye, Z., Deng, J., Ouyang, D., 2022. In silico formulation prediction of drug/cyclodextrin/polymer ternary complexes by machine learning and molecular modeling techniques. Carbohydr. Polym. 275, 118712 https://doi.org/10.1016/j.carbpol.2021.118712.

Ling, H., Luoma, J.T., Hilleman, D., 2013. A review of currently available fenofibrate and fenofibric acid formulations. Cardiol. Res. 4, 47. https://doi.org/10.4021/cr270w.

Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J., 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Deliv. Rev. 46, 3–26. https://doi.org/10.1016/s0169-409x(00)00129-0.

Lipton, Z.C., 2017. The mythos of model interpretability.

Lombardo, F., Berellini, G., Obach, R.S., 2018. Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 1352 drug compounds. Drug Metab. Dispos. 46, 1466–1477. https://doi.org/10.1124/dmd.118.082966.

Lou, H., Hageman, M.J., 2021. Machine learning attempts for predicting human subcutaneous bioavailability of monoclonal antibodies. Pharm. Res. 38, 451–460. https://doi.org/10.1007/s11095-021-03022-y.

Lou, J., Duan, H., Qin, Q., Teng, Z., Gan, F., Zhou, Xiaofang, Zhou, Xing, 2023. Advances in oral drug delivery systems: challenges and opportunities. Pharmaceutics 15, 484. https://doi.org/10.3390/pharmaceutics15020484.

Lundberg, S., Lee, S.I., 2017. A unified approach to interpreting model predictions.

Lyu, J., Wang, S., Balius, T.E., Singh, I., Levit, A., Moroz, Y.S., O'Meara, M.J., Che, T., Algaa, E., Tolmachova, K., Tolmachev, A.A., Shoichet, B.K., Roth, B.L., Irwin, J.J., 2019. Ultra-large library docking for discovering new chemotypes. Nature 566, 224–229. https://doi.org/10.1038/s41586-019-0917-9.

Mak, K.K., Pichika, M.R., 2019. Artificial intelligence in drug development: present status and future prospects. Drug Discov. Today 24, 773–780. https://doi.org/10.1016/j.drudis.2018.11.014.

Maltarollo, V.G., Gertrudes, J.C., Oliveira, P.R., Honorio, K.M., 2015. Applying machine learning techniques for ADME-Tox prediction: a review. Expert Opin. Drug Metab. Toxicol. 11, 259–271. https://doi.org/10.1517/17425255.2015.980814.

McKinney, W., 2010. Data structures for statistical computing in python. In: Proceedings of the Presented at the Python in Science Conference. Austin, Texas, pp. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., 2022. A survey on bias and fairness in machine learning. ACM Comput. Surv. 54, 1–35. https://doi.org/10.1145/3457607.

Mehta, C.H., Narayan, R., Nayak, U.Y., 2019. Computational modeling for formulation design. Drug Discov. Today 24, 781–788. https://doi.org/10.1016/j.drudis.2018.11.018.

Meng-Lund, H., Holm, T.P., Poso, A., Jorgensen, L., Rantanen, J., Grohganz, H., 2019. Exploring the chemical space for freeze-drying excipients. Int. J. Pharm. 566, 254–263. https://doi.org/10.1016/j.ijpharm.2019.05.065.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T., 2019. Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 220–229. https://doi.org/10.1145/3287560.3287596.

Molnar, C., Casalicchio, G., Bischl, B., 2020. Interpretable machine learning – a brief history, state-of-the-art and challenges. In: Koprinska, I., Kamp, M., Appice, A., Loglisci, C., Antonie, L., Zimmermann, A., Guidotti, R., Özgöbek, Ö., Ribeiro, R.P., Gavaldà, R., Gama, J., Adilova, L., Krishnamurthy, Y., Ferreira, P.M., Malerba, D., Medeiros, I., Ceci, M., Manco, G., Masciari, E., Ras, Z.W., Christen, P., Ntoutsi, E., Schubert, E., Zimek, A., Monreale, A., Biecek, P., Rinzivillo, S., Kille, B., Lommatzsch, A., Gulla, J.A. (Eds.), ECML PKDD 2020 Workshops, Communications in Computer and Information Science. Springer International Publishing, Cham, pp. 417–431. https://doi.org/10.1007/978-3-030-65965-3_28.

Montanari, F., Kuhnke, L., Ter Laak, A., Clevert, D.A., 2019. Modeling physico-chemical ADMET endpoints with multitask graph convolutional networks. Molecules 25, 44. https://doi.org/10.3390/molecules25010044.

Monteiro, N.R.C., Ribeiro, B., Arrais, J.P., 2021. Drug-target interaction prediction: end-to-end deep learning approach. IEEE/ACM Trans. Comput. Biol. Bioinform. 18, 2364–2374. https://doi.org/10.1109/TCBB.2020.2977335.

Mueller, E.A., Kovarik, J.M., van Bree, J.B., Tetzloff, W., Grevel, J., Kutz, K., 1994. Improved dose linearity of cyclosporine pharmacokinetics from a microemulsion formulation. Pharm. Res. 11, 301–304.

Musther, H., Olivares-Morales, A., Hatley, O.J.D., Liu, B., Rostami Hodjegan, A., 2014. Animal *versus* human oral drug bioavailability: do they correlate? Eur. J. Pharm. Sci. 57, 280–291. https://doi.org/10.1016/j.ejps.2013.08.018.

Nadkarni, P.M., Ohno-Machado, L., Chapman, W.W., 2011. Natural language processing: an introduction. J. Am. Med. Inform. Assoc. 18, 544–551. https://doi.org/10.1136/amiajnl-2011-000464.

Narayanan, H., Dingfelder, F., Condado Morales, I., Patel, B., Heding, K.E., Bjelke, J.R., Egebjerg, T., Butté, A., Sokolov, M., Lorenzen, N., Arosio, P., 2021. Design of biopharmaceutical formulations accelerated by machine learning. Mol. Pharm. 18, 3843–3853. https://doi.org/10.1021/acs.molpharmaceut.1c00469.

Netzeva, T.I., Worth, A.P., Aldenberg, T., Benigni, R., Cronin, M.T.D., Gramatica, P., Jaworska, J.S., Kahn, S., Klopman, G., Marchant, C.A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G.Y., Perkins, R., Roberts, D.W., Schultz, T.W., Stanton, D. T., van de Sandt, J.J.M., Tong, W., Veith, G., Yang, C., 2005. Current status of methods for defining the applicability domain of (Quantitative) structure-activity relationships: the report and recommendations of ECVAM workshop 52. Altern. Lab. Anim. 33, 155–173. https://doi.org/10.1177/026119290503300209.

Park, H., Otte, A., Park, K., 2022. Evolution of drug delivery systems: from 1950 to 2020 and beyond. J. Controlled Release 342, 53–65. https://doi.org/10.1016/j.jconrel.2021.12.030.

Parrott, N., Manevski, N., Olivares-Morales, A., 2022. Can we predict clinical pharmacokinetics of highly lipophilic compounds by integration of machine learning or *in vitro* data into physiologically based models? A feasibility study based on 12 development compounds. Mol. Pharm. 19, 3858–3868. https://doi.org/10.1021/acs.molpharmaceut.2c00350.

Paul, S., Baranwal, Y., Tseng, Y.C., 2021. An insight into predictive parameters of tablet capping by machine learning and multivariate tools. Int. J. Pharm. 599, 120439 https://doi.org/10.1016/j.ijpharm.2021.120439.

Petch, J., Di, S., Nelson, W., 2022. Opening the black box: the promise and limitations of explainable machine learning in cardiology. Can. J. Cardiol. 38, 204–213. https://doi.org/10.1016/j.cjca.2021.09.004.

Polli, J.E., 2008. *In vitro* studies are sometimes better than conventional human pharmacokinetic *in vivo* studies in assessing bioequivalence of immediate-release solid oral dosage forms. AAPS J. 10, 289–299. https://doi.org/10.1208/s12248-008-9027-6.

Poole, D.L., Mackworth, A.K., Goebel, R., 1998. Computational Intelligence: A Logical Approach. Oxford Univ. Press, New York Oxford.

Probst, P., Boulesteix, A.L., Bischl, B., 2019. Tunability: importance of hyperparameters of machine learning algorithms. J. Mach. Learn. Res. 20, 1934–1965.

Raschka, S., Kaufman, B., 2020. Machine learning and AI-based approaches for bioactive ligand discovery and GPCR-ligand recognition. Methods San Diego Calif 180, 89–110. https://doi.org/10.1016/j.ymeth.2020.06.016.

Raudys, S.J., Jain, A.K., 1991. Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans. Pattern Anal. Mach. Intell. 13, 252–264. https://doi.org/10.1109/34.75512.

Reinsel, D., Gantz, J., Rydning, J., 2018. The Digitization of the World from Edge to Core. IDC White Pap. 13.

Reppas, C., Kuentz, M., Bauer-Brandl, A., Carlert, S., Dallmann, A., Dietrich, S., Dressman, J., Ejskjaer, L., Frechen, S., Guidetti, M., Holm, R., Holzem, F.L., Karlsson, E., Kostewicz, E., Panbachi, S., Paulus, F., Senniksen, M.B., Stillhart, C., Turner, D.B., Vertzoni, M., Vrenken, P., Zöller, L., Griffin, B.T., O'Dwyer, P.J., 2023. Leveraging the use of *in vitro* and computational methods to support the development of enabling oral drug products: an InPharma commentary. Eur. J. Pharm. Sci. 106505 https://doi.org/10.1016/j.ejps.2023.106505.

Reymond, J.L., 2015. The chemical space project. Acc. Chem. Res. 48, 722–730. https://doi.org/10.1021/ar500432k.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. Why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Presented at the KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA. ACM, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778.

Rowe, R.C., Sheskey, P., Quinn, M., 2009. Handbook of Pharmaceutical Excipients. Libros Digitales-Pharmaceutical Press.

Ruddigkeit, L., van Deursen, R., Blum, L.C., Reymond, J.L., 2012. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. J. Chem. Inf. Model. 52, 2864–2875. https://doi.org/10.1021/ci300415d.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1, 206–215. https://doi.org/10.1038/s42256-019-0048-x.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., Zhong, C., 2022. Interpretable machine learning: fundamental principles and 10 grand challenges. Stat. Surv. 16 https://doi.org/10.1214/21-SS133.

Schaduangrat, N., Lampa, S., Simeon, S., Gleeson, M.P., Spjuth, O., Nantasenamat, C., 2020. Towards reproducible computational drug discovery. J. Cheminform. 12, 9. https://doi.org/10.1186/s13321-020-0408-x.

Schittenkopf, C., Deco, G., Brauer, W., 1997. Two strategies to avoid overfitting in feedforward networks. Neural. Netw. 10, 505–516. https://doi.org/10.1016/S0893-6080(96)00086-X.

Schneider, P., Walters, W.P., Plowright, A.T., Sieroka, N., Listgarten, J., Goodnow, R.A., Fisher, J., Jansen, J.M., Duca, J.S., Rush, T.S., Zentgraf, M., Hill, J.E., Krutoholow, E., Kohler, M., Blaney, J., Funatsu, K., Luebkemann, C., Schneider, G., 2020. Rethinking drug design in the artificial intelligence era. Nat. Rev. Drug Discov. 19, 353–364. https://doi.org/10.1038/s41573-019-0050-3.

Schwab, M., Karrenbach, N., Claerbout, J., 2000. Making scientific computations reproducible. Comput. Sci. Eng. 2, 61–67. https://doi.org/10.1109/5992.881708.

Sejdić, E., Lipsitz, L.A., 2013. Necessity of noise in physiology and medicine. Comput. Methods Programs Biomed. 111, 459–470. https://doi.org/10.1016/j.cmpb.2013.03.014.

Sinha, K., Murphy, E., Kumar, P., Springer, K.A., Ho, R., Nere, N.K., 2021. A novel computational approach coupled with machine learning to predict the extent of agglomeration in particulate processes. AAPS PharmSciTech 23, 18. https://doi.org/10.1208/s12249-021-02083-x.

Sotos, A.E.C., Vanhoof, S., Van Den Noortgate, W., Onghena, P., 2009. The transitivity misconception of Pearson's correlation coefficient. Stat. Educ. Res. J. 8, 33–55. https://doi.org/10.52041/serj.v8i2.394.

Stegemann, S., Leveiller, F., Franchi, D., de Jong, H., Lindén, H., 2007. When poor solubility becomes an issue: from early stage to proof of concept. Eur. J. Pharm. Sci. 31, 249–261. https://doi.org/10.1016/j.ejps.2007.05.110.

Steppe, J.M., Bauer, K.W., 1997. Feature saliency measures. Comput. Math. Appl. 33, 109–126. https://doi.org/10.1016/S0898-1221(97)00059-X.

Thite, N.G., Ghazvini, S., Wallace, N., Feldman, N., Calderon, C.P., Randolph, T.W., 2022. Machine learning analysis provides insight into mechanisms of protein particle formation inside containers during mechanical agitation. J. Pharm. Sci. 111, 2730–2744. https://doi.org/10.1016/j.xphs.2022.06.017.

Thomas, S., Palahnuk, H., Amini, H., Akseli, I., 2021. Data-smart machine learning methods for predicting composition-dependent Young's modulus of pharmaceutical compacts. Int. J. Pharm. 592, 120049 https://doi.org/10.1016/j.ijpharm.2020.120049.

Topliss, J.G., Costello, R.J., 1972. Change correlations in structure-activity studies using multiple regression analysis. J. Med. Chem. 15, 1066–1068. https://doi.org/10.1021/jm00280a017.

Topol, E.J., 2019. High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. 25, 44–56. https://doi.org/10.1038/s41591-018-0300-7.

Tosstorff, A., Menzen, T., Winter, G., 2020. Exploring chemical space for new substances to stabilize a therapeutic monoclonal antibody. J. Pharm. Sci. 109, 301–307. https://doi.org/10.1016/j.xphs.2019.10.057.

Trenfield, S.J., Awad, A., Goyanes, A., Gaisford, S., Basit, A.W., 2018. 3D printing pharmaceuticals: drug development to frontline care. Trends Pharmacol. Sci. 39, 440–451. https://doi.org/10.1016/j.tips.2018.02.006.

Vinarov, Z., Abrahamsson, B., Artursson, P., Batchelor, H., Berben, P., Bernkop-Schnürch, A., Butler, J., Ceulemans, J., Davies, N., Dupont, D., Flaten, G.E., Fotaki, N., Griffin, B.T., Jannin, V., Keemink, J., Kesisoglou, F., Koziolek, M., Kuentz, M., Mackie, A., Meléndez-Martínez, A.J., McAllister, M., Müllertz, A., O'Driscoll, C.M., Parrott, N., Paszkowska, J., Pavek, P., Porter, C.J.H., Reppas, C., Stillhart, C., Sugano, K., Toader, E., Valentová, K., Vertzoni, M., De Wildt, S.N., Wilson, C.G., Augustijns, P., 2021. Current challenges and future perspectives in oral absorption research: an opinion of the UNGAP network. Adv. Drug Deliv. Rev. 171, 289–331. https://doi.org/10.1016/j.addr.2021.02.001.

Vokinger, K.N., Feuerriegel, S., Kesselheim, A.S., 2021. Mitigating bias in machine learning for medicine. Commun. Med. 1, 25. https://doi.org/10.1038/s43856-021-00028-w.

Wang, W., Ye, Z., Gao, H., Ouyang, D., 2021. Computational pharmaceutics - a new paradigm of drug delivery. J. Controlled Release 338, 119–136. https://doi.org/10.1016/j.jconrel.2021.08.030.

Watson, D.S., Krutzinna, J., Bruce, I.N., Griffiths, C.E., McInnes, I.B., Barnes, M.R., Floridi, L., 2019. Clinical applications of machine learning algorithms: beyond the black box. BMJ l886. https://doi.org/10.1136/bmj.l886.

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018. https://doi.org/10.1038/sdata.2016.18.

Yu, L.X., Amidon, G., Khan, M.A., Hoag, S.W., Polli, J., Raju, G.K., Woodcock, J., 2014. Understanding pharmaceutical quality by design. AAPS J. 16, 771–783. https://doi.org/10.1208/s12248-014-9598-3.

Yu, L.X., Raw, A., Wu, L., Capacci-Daniel, C., Zhang, Y., Rosencrance, S., 2019. FDA's new pharmaceutical quality initiative: knowledge-aided assessment & structured applications. Int. J. Pharm. X 1, 100010. https://doi.org/10.1016/j.ijpx.2019.100010.