View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by LJMU Research Online

Characterisation of data resources for *in silico* modelling: benchmark datasets for ADME properties.

Article highlights:

- Over 140 datasets covering 24 ADME properties were identified from the literature
- Characteristics influencing the modelability of available data were investigated
- 31 benchmark datasets were assessed according to defined criteria for modelability
- Recommendations are provided for publishing and assessing suitability for modelling

Abstract

Introduction: The cost of *in vivo* and *in vitro* screening of ADME properties of compounds has motivated efforts to develop a range of *in silico* models. At the heart of the development of any computational model are the data; high quality data are essential for developing robust and accurate models. The characteristics of a dataset, such as its availability, size, format and type of chemical identifiers used, influence the modelability of the data.

Areas covered: This review explores the usefulness of publicly available ADME datasets for researchers to use in the development of predictive models. More than 140 ADME datasets were collated from publicly available resources and the modelability of 31selected datasets were assessed using specific criteria derived in this study.

Expert opinion: Publicly available datasets differ significantly in information content and presentation. From a modelling perspective, datasets should be of adequate size, available in a user-friendly format with all chemical structures associated with one or more chemical identifiers suitable for automated processing (e.g. CAS number, SMILES string or InChIKey). Recommendations for assessing dataset suitability for modelling and publishing data in an appropriate format are discussed.

Keywords: ADME, *in silico*, database, dataset, modelling, modelability

1. Introduction

The ability of a compound to elicit any biological effect (be that a desirable therapeutic effect or an undesirable toxic effect) depends on two factors: the intrinsic activity of the compound and its potential to reach the site of action in sufficient concentration for the requisite time period. The likelihood of the compound reaching any given site depends on both the external and internal exposure [1]. External exposure relates to the possibility of the compound reaching an organism, i.e. the concentration present in the environment to which the organism is exposed. Internal exposure relates to the uptake and distribution of the compound to sites of action within an organism, its metabolism and ultimate elimination from the body. These properties of a compound, termed absorption, distribution, metabolism and excretion (ADME), will determine the concentration-time profile of the compound at a given site of interest or in the body as a whole. These factors are therefore important in determining the *in vivo* response to a potential therapeutic agent or a toxicant as they may have a significant modulating effect on overall activity. Knowledge of ADME properties is critical in the development of new drugs and in assessing the risk posed by compounds such as industrial chemicals, food additives, pesticides and environmental pollutants, to which an organism may be exposed to [2]. For example, negligible absorption via a given route may provide a scientific justification for waiving certain toxicological studies. ADME data can also assist in the interpretation of animal studies, particularly with respect to identifying inter-species differences or similarities [3]. Moreover, information on ADME properties is crucial for the development of appropriate toxicity testing strategies, including the selection of appropriate species and doses for toxicity testing, and for risk assessment enabling comparison of internal dose in experimental animals and humans [4].

Knowledge of ADME properties of category members is also recognised as being useful in substantiating read-across approaches for data gap filling [5]. ADME data can provide insight

as to which tissues/organs are exposed, as well as the kinetics of this exposure. Within a well-defined category, or for structural analogues, that are assumed to be acting by the same mechanism of action, knowledge of toxicokinetics will assist in justifying a read-across prediction.

ADME data can be generated using *in vivo*, *in vitro* and/or *in silico* methodologies. However, *in vivo* and *in vitro* evaluation methods are costly and time consuming; with *in vivo* methods having the additional, undesirable requirement for animal use. These methods are not suitable for testing large numbers of chemicals. Therefore, over the past three decades interest has grown in *in silico* methodologies and consequently a large number of models have been developed to evaluate ADME properties, particularly as part of the drug discovery process. The purpose of the present paper is not to review the models themselves as several excellent reviews regarding the development of such models have already been published [2, 6-11]. The aim here was to evaluate the nature of the publicly available datasets for ADME properties, such as to inform potential, future model builders as to the suitability for modelling of each dataset. The advantage of developing *in silico* models is that, being high-throughput and low cost, they can be used to (virtually) screen thousands of compounds, in a short time to help prioritise compounds for further testing or development [11].

To develop a reliable *in silico* model, ideally a significant number of high quality experimental data are needed [9]. Whilst screening for ADME properties may be routine in drug development, this does not result in a commensurate high availability of data in the public domain that may be used by modellers - for reasons of confidentiality [12]. Of those data that are available, they may be stored within a database and/or a dataset. The former is defined here as an organised collection of data that may be logically searched or retrieved, under the control of database management software; the latter refers to any set of data, usually presented in table format, that may be readily processed computationally (the dataset

may be extracted from a database, or available in a standalone format). Many businesses will possess their own in-house ADME databases; however, publicly available data that can be obtained rapidly and with little or no cost are preferable, particularly for academia. There are a number of publicly available databases, such as PubChem [13], DrugBank [14], ChEMBL [15] and ACToR [16], where a reasonable number of ADME–related data can be sourced. However, such data for ADME endpoints very often need to be extracted using appropriate data retrieval techniques [17] and then collated and organised in a tabular format. A relatively large number of ADME data are available in the ADME SARfari database [18, 19]. However, data for all ADME endpoints are collated together in one file, which necessitates selective extraction of the data of interest. An additional difficulty is that the chemical identifiers are stored in a separate file, therefore the chemical structures from one file have to be assigned to datapoints from another file. ADME-related data, compiled in a single place to create a tabular data matrix, are more attractive and useful for modellers. Ready-made ADME datasets are usually obtained from peer-reviewed literature, reports or from ADME online databases, such as the PharmacoKinetics Knowledge Base (PKKB) [20].

Whilst many datasets are available, there is no standard template, generally accepted by scientific journals, for publishing data i.e. no formally adopted standards on how chemical structures should be represented, which biological information should be available and the level of experimental detail that should be recorded and in which format. There are examples of published guidelines, such as the reporting guidelines for bioactive entities— the Minimum Information About a Bioactive Entity (MIABE) — which is a formal list of the items of information about a molecule, its properties, production, physicochemical characteristics, biological activities (obtained using *in vitro* cell-free assays, cellular assays, whole-organism studies or pharmacokinetic studies) that are recommended to be published by the data provider [21]. These guidelines were developed by representatives of

pharmaceutical companies, data resource providers and academic groups and are proposed to be used during manuscript preparation to make all data on bioactive molecules readily available in a single common format, with a full and consistent data description. Fourches et al. proposed a set of procedures for chemical data curation that should be followed at the onset of any molecular modelling investigation [22]. Additionally, for certain endpoints guidelines or checklists for recording experimental information have been promoted previously [23, 24]; recording and analysis of metadata and associated data governance considerations (in the context of different toxicity endpoints) has been reviewed by Fu et al [25]. However, none of these guidelines or strategies have been implemented in a standard procedure for publishing data in publicly available domains. Therefore, publicly available datasets differ in terms of information content and presentation as well as in overall intrinsic data quality. Modellers using readily-available datasets from public resources rely largely on the data providers for curation, accuracy and consideration of quality. Whilst any dataset compilation should be quality checked for the correctness of the chemical and biological information and overall intrinsic quality of the data, these processes are not obligatory and are seldom, if ever, recorded.

Quality assurance of published datasets should consist of three stages:

(i) Chemical structure characterisation

It is crucial to ensure that the chemical structures are reported correctly and characterised by consistent, unambiguous chemical identifiers. Incorrect structures [26], inconsistency between the chemical identifiers [27] and/or their ambiguity (i.e. whether an identifier matches more than one chemical structure) [28] pose a significant problem for the development of predictive models. Aside from the correctness of chemical identifiers, the curation of chemical structures influences the predictive performance of models significantly

[22]. The data curation process may include: the removal of mixtures, inorganics and organometallics; structure conversion; cleaning / removal of salts; normalisation of specific chemotypes; treatment of tautomeric forms; and identification / removal of duplicates. [22]. Despite automation, structure curation may still require significant manual inspection both to determine which changes are needed in the analysis stage and post processing to verify that the changes have the desired effect. Only well curated chemical data have the potential to generate models with reasonable robustness and reliability [22, 29].

(ii) Biological information assessment

Biological details, such as the nature of the experiment, whether it is *in vivo* or *in vitro*, species, sex, strain of animals /organisms and experimental conditions (duration, dose, vehicle etc.) should be carefully assessed for consistency and correctly reported. For example, in terms of developing predictive models, the vehicle (or co-administered substances) used for a given experiment may have a significant effect on the outcome. Modellers need to determine in which cases results using different vehicles can be combined and when they require the development of alternate models. There are a number of studies, both experimental and *in silico* that have investigated the influence of the vehicle on biological activity. For example, Jowsey et al. showed that for certain vehicles the variation in the measured skin-sensitisation potential of compounds (determined using the local lymph node assay (LLNA)) was no greater than the expected experimental variability. However, for other vehicles the potency values recorded for the same compound in different vehicles varied by a factor of 10 [30]. Mistry et al. produced random forest and decision tree models to investigate the influence of different drug-vehicle combinations on drug toxicity, enabling vehicle selection to be optimised to reduce toxic effects [31]. Ghafourian et al. developed quantitative structure activity relationships (QSARs) models to predict skin permeability coefficients of compounds dissolved in different vehicles [32]. Moreover, when data are

combined from more than one species, it can have a great effect on the results. Significant inter-species differences in response will affect the accuracy of predictions when scaling between one species and another [24]. Also intra-species variations, such as differences in age, gender can influence the *in vivo* outcomes and subsequently the correctness of predictions [24]. Such issues of variability can only be determined if the relevant biological information is recorded.

When data are collated from different sources, multiple values of biological activity for the same chemical may be found even within one dataset. Where the values are comparable, it is relatively easy to define a representative value using standard statistical procedures, such as median [33], geometric mean [34, 35] or arithmetic mean [36]. When values for one chemical differ from others by orders of magnitude or are outside of the range of ±50% of the median, such data can be considered as outliers and may be omitted [37]. Another option is to remove any obviously unreliable data points and then select the lowest value for the EC50 from the remaining data [38]. This is a conservative approach where the lowest dose associated with a given toxic response is assumed is used. It is worth mentioning, that the presence of multiple, comparable values for the same chemical can increase confidence in the data and this may be expressed as a confidence score (CS), and the use of such data will consequently improves the robustness of the model [39].

In cases where conflicting biological results have been identified, more detailed investigation into the studies to identify potential sources of variation is required [40]. On this stage, the importance of understanding biological data is vital. Where an obvious reason for discrepancy in biological activity cannot be found then a weight of evidence approach may be used i.e. confirmation is sought from additional sources. If the conflict between data values cannot be resolved then the data should be excluded from modelling.

(iii) Assessment of intrinsic data quality

Assuring data quality, although time consuming, is of key importance as it directly influences the predictive performance of models. There is a universal agreement that models are only as good as the data on which they are based [41]. The importance of data quality assessment and the need for an appropriate scheme and/or tool for such assessments is widely recognised and has been discussed in detail elsewhere [40, 42-45]. Ideally, the data should be reported in a manner which is conducive to quality assessment [46]. This requirement is particularly pertinent to the drug discovery process, here the emphasis has shifted from the development of predictive models, to the need for techniques to manage, curate, and integrate large amounts of potentially useful data [46].

Such quality controlled datasets are at the heart of developing reliable *in silico* models. There are various methods for generating models to predict biological endpoints, such as statistical techniques developing e.g. QSARs models or expert system approaches such as structural alerts. These methodologies have been discussed in depth elsewhere [47] and they are beyond the scope of this review as here the focus is on the data itself rather than the models generated. When selecting an *in silico* model to use it is crucial to understand the strengths and limitations of each model [48]. The predictive strength of an *in silico* ADME model influences at which stage of the drug development process the model can be most gainfully employed [48].

For modellers, there are additional characteristics of a dataset that significantly influence the ability to model the data therein (referred to in this paper as "modelability") [49]. In general, the most favourable datasets for modelling consist of a large number of compounds characterised by machine-readable identifiers (enabling automated processing) and recorded in a user-friendly format. These aspects of datasets that influence their modelability will be

discussed and evaluated in this review. Therefore, the aims of this work were: to define the characteristics of datasets that are important for modellers to consider when selecting the most appropriate dataset for model development (i.e. to determine characteristics influencing modelability); to identify publicly available resources for ADME properties; to assess the usefulness of these ADME datasets for modelling purposes; and, finally, to make recommendations to help modellers select the most appropriate dataset(s).

2. Characteristics of datasets that influence their modelability

In general, datasets appropriate for modelling for any endpoint/biological activity combine two types of information: chemical and biological. The manner in which such data are presented to the user constitutes the characteristics of the dataset and determine their modelability. The relationships between these factors are shown in Figure 1 and are discussed individually below.

The information needs to be presented in a well organised and transparent manner which is easy to understand and use. Specific characteristics that influence suitability for modelling include: availability, size, format and types of chemical identifiers used (which determines how quickly and easily the data can be processed computationally) and accuracy. Whilst intrinsic data quality is essential for building reliable models, data quality assessment has been extensively reviewed elsewhere, as indicated above. Hence this review relates to the modelability of data, in terms of its presentation to the user, rather than considering intrinsic data quality.

2.1 Availability

With regard to the availability of data, there are public, commercial or in-house databases. Commercial databases may be richer in data and more diverse than publicly available

sources. In-house databases may be more diverse or may be specifically focused on the chemical spaces of interest to the company, however, they have the significant advantage that the data are often more consistent with less experimental variability than other sources [50]. Publicly available datasets may be smaller; larger datasets are generally collations from multiple sources and therefore have much higher experimental variability due to inter-laboratory differences in procedures. The main advantage of public datasets is their free accessibility; hence they are highly desirable and extensively used by modellers.

2.2 Size

Another aspect important for the development of a model is the size of the dataset, which has a significant impact on chemical space and range of activity covered by the model. Publicly available datasets are often created by collation of data from various sources. This means that the data may be measured in different laboratories using different experimental procedures/conditions resulting in significant inter-laboratory differences [12]. Generic and endpoint-specific sources of experimental variation are discussed in Madden et al. [24]. As smaller datasets may be iteratively collated into larger datasets, there is a possibility of repetition of data, hence the number of unique compounds needs to be confirmed. Collation may lead to problems of perpetuating errors appearing in earlier publications or introducing new errors on processing. Such heterogeneous datasets cover a wider chemical space, this diversity renders them attractive to modellers. However, variability may reduce the maximal accuracy that can be obtained for the models [51, 52]. For collated datasets, it is important that references to the original sources of experimental results are provided allowing for verification of results, if necessary, and giving access to additional information for the chemicals tested. Conversely, homogenous datasets use data from a single study (or a small number of studies) carried out under the same experimental protocol, in the same laboratory,

possibly by the same operator. Such datasets are smaller, have reduced experimental variability and may cover a narrower activity range, resulting in models with potentially greater accuracy but with a reduced applicability domain.

2.3 Format

Data can be recorded in many different file formats, examples include: Excel, comma-separated values (CSV), Text, Word, portable document format (pdf), structure-data file (sdf), hypertext markup language (html) and a number of other formats. It is important for modellers to have data in a user–friendly format, which is easy to handle and allows for ready data analysis and/or manipulation. An Excel file, whilst limited in some aspects (e.g. linkage to chemistry), is an appropriate format as it allows for a variety of manipulations, as well as easily importing or exporting data (both individual values and large collations). Data recorded in CSV and Text formats are also suitable for modellers, as these data can be easily imported to Excel files. The sdf format offers clear visualisation of chemical structures [53], however, not all software recognise this format and it is not possible to readily manipulate data in this format. Word and pdf formats are less attractive to modellers, as they cannot be processed by computational chemistry or statistical software. Some datasets have been published graphically, e.g. in graphic interchange format (gif) however, again this is not suitable for modelling as data must be manually converted into a more suitable format.

2.4 Types of chemical identifiers

Finally, the type of chemical identifiers used has a significant impact on model development. There are a number of potential identifiers and systems by which a chemical structure can be represented [40]:

- Nomenclature: International Union of Pure and Applied Chemistry (IUPAC) names, common names, inventory or database names
- Unique Identifiers: Chemical Abstracts Service (CAS) registry numbers, inventory or database numbers (e.g. ChemSpider ID number; European INventory of Existing Commercial chemical Substances (EINECS) number)
- Line notation methods: Simplified Molecular Input Line Entry System (SMILES) strings [54], IUPAC International Chemical Identifier (InChI) code [55], empirical formula
- Pictorial representations in 2 or 3 dimensions.
- Representation of 3-dimensional structures (e.g. mol, sdf files)

To ensure the accuracy of chemical identity, at least two, and ideally three concordant identifiers should be provided in the dataset. Having three concordant identifiers, such as chemical name, CAS registry number and structure, is a strong indication that the chemical identity is correct. Any lack of consistency between the chemical identifiers for a compound would require further investigation or exclusion of the compound from the dataset.

From the modeller's perspective, it is preferential to utilise identifiers encoding the chemical structure in a form which is easy to interpret and can be easily input into, and recognised by, a range of different software. SMILES strings fulfil these criteria; therefore, datasets with chemicals identified by SMILES annotations are valuable for developers of predictive models. However, SMILES strings also have their shortcomings; they focus on molecules with bonds that fit the 2-electron valence model, have a limited array of stereochemistry types and have no standard for handling aromaticity [56]. Also the issue of multiple tautomers of the same molecule cannot be handled adequately by SMILES notation as it stretches the conceptual limit of a unique chemical identifier [57, 58]. Moreover, there is no standard means to generate canonical SMILES, as toolkits create canonical SMILES differently [56]. There have been attempts to overcome this limitation, such as the

introduction of universal SMILES, based on the InChI canonicalisation algorithm [56], but this remains a restriction to their universal applicability.

These four characteristics of datasets, (availability, size, format and types of identifier used), were considered to be key criteria for assessing their suitability for modelling and were adopted as the evaluation criteria for the datasets investigated within this study.

3. The identification of publicly available ADME datasets

In general, ADME properties can be classified into one of two categories: physiological and physico-chemical [50]. A number of physiological ADME properties have been identified in this investigation covering the four fundamental processes of absorption, distribution, metabolism and excretion. Physiological ADME properties can be further sub-divided into *in vitro* ADME properties (such as uptake across the human epithelial colorectal adenocarcinoma (Caco-2) cell line or the Madin-Darby Canine Kidney (MDCK) cell line) and *in vivo* pharmacokinetic properties (such as oral bioavailability, human intestinal absorption (HIA), plasma protein binding (PPB) and urinary excretion). Table 1 summarises, with brief definitions, the physiological ADME properties investigated in this study. Table 1 is not exhaustive, but covers a wide range of the most significant pharmacokinetic / toxicokinetic properties for which models are required. For the purposes of this investigation some related properties have been grouped together (refer to third column in Table 1) as the individual properties are representative of related (or the same) endpoints. For example, datasets for percentage absorption (% abs) – here referring to absorption across the gastro-intestinal tract have been grouped together with datasets that refer specifically to percentage human intestinal absorption (%HIA). Similarly, datasets for fraction bound/fraction unbound in plasma were grouped with percentage plasma protein binding (%PPB) as the latter datasets were more populous. Physico-chemical properties (governed by the Laws of Physical

Chemistry) include aqueous solubility, the logarithm of octanol–water partition coefficient (log P), logarithm of octanol–water distribution coefficient (log D) and acid dissociation constant pKa [50]. These are data rich properties that have been extensively reviewed elsewhere, [59-63], thus they were excluded from this analysis.

Data for the ADME properties identified in Table 1 were retrieved by a thorough literature search and review of online resources. Only freely available datasets were considered as they are most useful for modelling. Two online ADME databases: PKKB [18] and PK/DB [64] were used as the starting point for this search. These databases provide compiled datasets for several ADME endpoints, including: Caco2 permeability, blood-brain permeability, PgP inhibition, oral absorption and oral bioavailability. In addition to the online databases, two extensive reviews were used as a source of available ADME datasets [1, 8]. By preference, the most recent collations of the experimental data for the ADME properties identified were usually considered as they were the most comprehensive and up to date.

In total, 141 ADME datasets were identified using the above procedure, these covered the majority of the individual (or grouped) ADME properties given in Table 1. Information for each of these endpoints were stored on separate Excel spreadsheets that included the following fields: dataset identity number, key reference source, availability/licencing requirements, format of data (e.g. sdf, pdf etc.), number of compounds in dataset, chemical identifiers used (e.g. names, SMILES, CAS), nature of biological data (i.e. *in vivo* or *in vitro* assay; species used), additional endpoint data available from the same references (some references included large collations of multiple endpoints and / or physico-chemical property data), diversity of chemicals (i.e. whether all chemicals were drugs, drug-like or representative of many areas of chemical space), whether the data were a collation or obtained from a single source, availability of supporting documentation (i.e. availability of original references), additional comments (where appropriate) and hyperlinks to the original

publication / website from where the dataset can be obtained (subject to appropriate user / institutional journal access rights). This Excel file is available as supporting information.

Some ADME endpoints, such as: absorption, bioavailability and blood-brain barrier (BBB) partitioning are rich in data, for others, there is a scarcity of data. For three endpoints (uptake by the respiratory tract, fraction bound/unbound in tissues and kinetic data) no suitable datasets were obtained. However, for uptake by the respiratory tract studies on single chemicals have been identified and for kinetic data (e.g. Cmax, Tmax) values for a small number of individual compounds are available in DrugBank [14]. From inspection of the datasets different recording formats were apparent. Approximately 70 of the datasets were published as pdf files and only 33 datasets were in Excel format (less common formats used were Word files, sdf files, CSV etc.). The datasets ranged in size from fewer than ten compounds to several thousand compounds. The majority of datasets were collations from diverse sources using different experimental procedures, usually, the source of the primary experimental data was given (level of experimental detail was variable). Only ten datasets provided data from a single source where all chemicals were tested according to the same experimental protocol. The chemicals were characterised using up to three identifiers, including: name, CAS number, SMILES string and/or pictorial representation; however, in most cases, the compounds had only one identifier. Since the majority of ADME data came from drug discovery, they were usually identified by common/marketing names, which may present problems for obtaining alternative identifiers that encode structure, such as SMILES strings or InChI keys. Although not an exhaustive list of available datasets, not least as the number is continually expanding, this was considered a good selection to carry forward to the next step of the process -assessment of suitability for modelling.

4. The assessment of the selected datasets in terms of their usefulness for modelling purposes

From the modeller's perspective, the "ideal" dataset should be publicly available (in a user-friendly file format such as Excel), reasonably large with diverse chemicals and described by an adequate number of concordant identifiers. A selection of the ADME datasets identified were assessed for their suitability for modelling using the four characteristics discussed above (i.e. availability, size, format and type of chemical identifiers used). In general, the largest data set for each ADME endpoint was investigated. In some cases, more than one dataset was evaluated e.g. where datasets were of a similar size, or where datasets covered different biological effects / activities (such as different sub-types of transporters). In total, 31 datasets were selected for assessment. Table 2 shows the datasets chosen for assessment together with information on their original format and available chemical identifiers.

4.1 Availability

The free availability of data was the essential criterion when considering ADME datasets in this analysis. Other sources of ADME data, such as commercial databases, were not considered. Thus, all ADME datasets identified herein are publicly available either from publications (or supporting information / on request from authors), book chapters or online databases. Certain online databases require registration with the database administrator prior to receiving a free licence or password, journal access is subject to usual user access restrictions.

4.2 Size

The size of the dataset (i.e. the number of unique compounds) was chosen as the next selection criterion for the assessment of their usefulness for *in silico* modelling. The optimal size of a dataset for modelling purposes is arbitrary; it depends on several factors, such as the

(statistical) method used to generate the model [29]. A larger dataset often means greater diversity, and hence coverage, of chemical space which, in turn, provides a wider applicability domain for any model developed therefrom.

Moreover, models built upon large training datasets are likely to be more robust than those developed using smaller training sets. In predicting toxicity, ensuring consistency in the putative mechanism of action of compounds in the training set (which may lead to reduction in training set size) is an important consideration, this is, however, less of an issue in predicting ADME endpoints [65, 66].

Analysis of the selected datasets shows that they range in size from fewer than ten compounds to thousands of chemicals. The smallest datasets were those for compounds crossing the blood:testis barrier with only 6, 7 and 10 chemicals in the three available datasets, respectively [67-69]. The largest datasets, comprising several thousand compounds, were obtained for: interactions with transporters (e.g. 3,763 chemicals interacting with 12 membrane transport proteins [70]), BBB partitioning (2,053 chemicals with binary data [71]), or Caco2 permeability (1,301 compounds with the maximum and minimum values of permeability given [72]). In some cases, it may be appropriate to sub-divide the dataset into smaller groups. For example, sub-dividing acidic and basic compounds, or those that interact with specific transporters, however, such decisions remain within the remit of the modeller and the purpose for which they are using the dataset

4.3 Format

The datasets selected for analysis were published in a number of different formats (see Table 2): Excel (15 datasets), .pdf (10 datasets), .sdf (2 datasets), text (1 dataset), Word (2 datasets) and .gif (1 dataset). If the data are to be used for modelling, they should be recorded in a format which is easily accessible and allows for quick and simple data extraction and

analysis. Although proprietary, the Excel format is user-friendly for reporting and storing data and is widely available. Excel spreadsheets allow for transparent and logical storage of multiple data points which can be readily manipulated; a single datum point or, group of data, can be extracted from Excel and processed using a variety of other software. Therefore, the existing 15 datasets published in Excel format were assessed as having a user-friendly data format. It should be appreciated that whilst good for storing data, Excel is not a database format and is not able to link to chemistry-based searches etc.

The usability of other data formats was assessed by attempting to convert the data into a format that could be readily incorporated into Excel. Conversion from Text format (as applied to a BBB dataset [71]), using appropriate delimiters in Excel, was the most straightforward conversion. Bioavailability and transporter datasets, reported as Sdf files [73, 74], could be readily converted to CSV formats and subsequently read into Excel, therefore were also considered suitable formats. Conversion from Word files (as performed for blood : testis and milk : plasma datasets [67, 75]) is easily achieved for some formats (e.g. if data are stored in a standard table). However, where there is additional formatting within the table (such as merged cells) then manual inspection of the transferred data is necessary – hence this format is only suitable for smaller datasets. Conversion from pdf to Excel (using Adobe Acrobat Pro software [76] resulted in several problems such as: specific characters being incorrectly converted, columns being inappropriately merged, data points appearing in incorrect cells etc. This means that careful, time-consuming manual curation is essential, rendering pdf a less suitable format. The dataset for skin absorption (Jmax) values was published as picture format, .gif file [77]. As all data then requires manual extraction this was considered the least useful format for the datasets considered here.

4.4 Chemical identifiers

Unambiguous, chemical identifiers that can be used to relate a chemical structure to a given property are essential for modelling. There are well recognised problems with using chemical names and CAS numbers (e.g. lack of consistency, conversion errors on entering into Excel etc. [70, 78,]), although they should add certainty to the identity of chemicals. Many structures have historically been stored as SMILES strings, although these can have different formats and interpretations and isomers can also present problems. The InChI format may be a better solution, but even that does not assist in the capture of 3-D information - that would require, for example, an .sdf file. The InChI uses a layered format to represent all the available structural information (formula, connectivity, isotopes, stereochemistry and tautomers) relevant to compound identity [79].

The selected datasets differed in the number of identifiers used: for only three datasets chemicals were characterised by three identifiers: name, CAS number and SMILES strings. Five datasets were characterised by name and CAS number; nine datasets by name and SMILES strings; 14 datasets provided the name only. Machine-readable identifiers (such as SMILES strings) are important to modellers to enable automated processing of the data in a range of software. Translating other identifiers into SMILES can be achieved by automated processing, however there are caveats to this. ChemSpider [80] can be used to convert from names to InChIKeys, from which an MDL molfile can be downloaded. The OpenBabel [81] node within KNIME [82] can then convert MDL to SMILES. Problems arise where non-systematic (e.g. generic or brand names) or incorrect names are used and it should be noted that some structures may not be available within public data resources or may be incorrectly recorded. Manually obtaining and/or curating identifiers may therefore be necessary to ensure accuracy. Datasets comprising two or more concordant identifiers, including at least one in a machine-readable format are therefore considered most suitable for modelling. On the other hand, datasets with only one identifier, most often name, are the least favoured for modellers,

as the translation of names into an identifier encoding structure, such as SMILES, is required. Additionally, the presence of only one identifier for the chemical adds uncertainty to the identity of the tested compound, especially if the names are incorrectly recorded or non-standard. The summary outcome of assessment of the datasets considered, in terms of availability, size, format and chemical identifiers used, is given in Table 2.

5. Conclusions

Publicly available datasets have been sourced for 24 ADME properties. However, only a small number of endpoints, such as oral absorption and bioavailability are rich in experimental data; for many other ADME properties, there is a scarcity of data. The data are the bedrock of the development of computational models and are of key importance in drug development and / or to assess the potential for toxicity. The so-called modelability of a dataset is influenced by its availability, size, format and types of chemical identifier used. These criteria were used to assess the usefulness of the datasets obtained in this study for modelling purposes. Generally, the largest datasets were chosen for assessment for each ADME endpoint, as larger datasets are favoured by modellers, although those considered varied in size from fewer than 10 compounds to several thousand. With regards to format, approximately half were published in Excel (considered to be the most favourable format) with the remaining datasets being in sdf, pdf, Text, Word or Gif formats. SMILES strings were the preferred identifier, although these were only available for twelve of the datasets; SMILES strings were created for the remaining datasets.

In summary, this study has developed criteria for assessing the usefulness of publicly available datasets for (Q)SAR modelling and applied these criteria to 24 ADME endpoints. As a result, 31 “benchmark” ADME datasets are identified for modellers to use (details are available within the supplementary information). 30 of these datasets are provided in Excel

files as supplementary information. One dataset for skin absorption (Jmax) values [77], was published as gif file, this was assessed as unsuitable for modelling and consequently has not been converted into an Excel file format here.

6. Expert opinion

Although, there is still a scarcity of experimental data for some ADME properties, especially in public resources, a large number of ready-made datasets have been published for other endpoints. Such datasets present great potential for generating *in silico* models. However, the datasets differ in their characteristics as there are no standard widely accepted guidelines as to how they should be reported. Here we undertook an assessment of usefulness of publicly available ADME datasets for modelling and have shown that adherence to standardised guidelines for format and contents would be of great assistance for modellers.

The choice of dataset is in part determined by the type of model being developed, for example considering whether a global model (requiring a larger, more chemically diverse dataset) or a local model is more desirable. It is acknowledged that, where appropriate, consideration must also be given to mechanism of action, although this is generally more relevant to toxicity prediction than ADME prediction. Intuitively modellers may select the largest dataset, leaving other factors, such as format and types of chemical identifiers used, as second-tier criteria for selection. This study has shown data format has a significant impact on dataset modelability. Therefore, size and format should both be considered pragmatically in the first stage of dataset selection. Also important, is the correct identification of chemicals by ideally three concordant identifiers. Datasets with chemicals identified by name only are potentially the most misleading and require additional processing i.e. names require transcription into identifiers encoding structure. Although this process may be automated to some extent, manual verification or and / or retrieval of identifiers may still be necessary.

Misspelt, ambiguous or non-standard names can be difficult to verify and therefore such chemicals have to be removed from the dataset. Therefore the presence of at least one identifier which encodes the chemical structure and is machine-readable is preferable for modelling purposes. SMILES format is the most popular line notation, being easily generated and interpreted by both humans and software. However, there are disadvantages to SMILES strings, such as a limited array of stereochemistry types, lack of standardisation for representing the aromaticity or handling multiple tautomers. The greatest limitation of the SMILES format is that there is no standard means to generate a canonical representation [56]. Therefore, it is recommended to standardise SMILES strings using a single canonicalisation algorithm for the dataset of interest. Recently an alternative line notation identifier (the InChI string) has become more widely applied. Although the InChI does resolve some of the issues with SMILES it has the disadvantage of requiring software for its interpretation [99, 100]. Aspects associated with the intrinsic accuracy of the data, such as correctness of chemical and biological details, are usually assessed by the developers of the datasets. Although a modeller may use the data without further quality assessment, use of data that has been quality assured in some way is preferable. To enable checking of chemical and / or biological information within a dataset, the availability of the reference to the original study is crucial. Therefore, the providers of the datasets should ensure this information is recorded, wherever practicable, during the collation and organisation of the data. In the case of multiple values being recorded for the same chemical the modeller must judge the most pragmatic approach; for example using a median or mean if the values are sufficiently "similar", or rejecting as outliers values that are extreme or anomalous. For this reason modellers require some appreciation of the level of variation inherent within a given assay to enable realistic boundaries for results to be determined. This requires information on how the result was obtained, the experimental protocol used and how experimental factors can impact the test

result (e.g. solubility limitations for compounds studies, use of solvents, alternative routes of exposure, etc.) [101]. Weight of evidence approaches may assist in cases where highly variable or contradictory data are recorded, resulting in higher confidence being assigned to a particular value or identifying data that should be excluded. The modelling process itself may identify anomalous data points which can then be queried or rejected as necessary. The overall characteristics of an ideal dataset and the chemical and biological information on which these are based are given in Table 3.

As assessment of inherent data quality has been dealt with in other reviews such assessments are not reported here, rather the reader is referred to key references in the area [40, 42-45]. The recommendations given below refer to assessment of dataset suitability for modelling in terms of coverage of chemical space, the presentation of the data and its accessibility to modellers.

6.1 Recommendations to help modellers select the most appropriate dataset

Based on the outcomes of the above assessment of the usefulness of the selected ADME datasets for modelling purposes, recommendations for selecting the most appropriate dataset are given below:

1. The availability, size, format and types of chemical identifiers used should be considered, pragmatically during selection of the dataset for *in silico* modelling purposes.
2. Excel remains a useful format for recording datasets for modelling, while CSV and text files are suitable alternatives. The pdf file format should only be used if data are not available in any other format; gif format requires manual re-entry of all data and is therefore least useful and not recommended for publishing datasets.
3. Datasets with concordant multiple chemical identifiers, with at least one encoding the chemical structure in a machine-readable format (e.g. SMILES string), are preferred.

Datasets with only chemical names are potentially the most error-prone; names may be ambiguous or misspelt. The use of such datasets should be avoided wherever possible.

4. Any conversion of a dataset (e.g. from pdf to Excel, or CAS number to SMILES strings) requires subsequent manual inspection to confirm that the data have been transferred / transformed correctly. Special attention needs to be paid to specific characters, symbols, missing data etc.

Finally, it would be very useful to encourage data providers to report / publish their datasets in modelling-ready format adhering to the principles outlined above. The "benchmark" datasets provided here (supplementary information) can serve as examples. Having datasets in such a format will help modellers reduce not only time and effort in developing *in silico* models, but also will help to reduce transcription errors, which may be made during subsequent data archiving processes. The investigation by Young et al. showed that data translation errors range from 0.1 to 3.4% depending on the database in question [26]. Additionally, Fourches et al. showed that these transcription errors can complicate the generation of QSAR models [22]. Therefore, it is highly recommended to develop and apply standard guidance how to report datasets, especially in public resources. Creation of a centralised repository for the datasets in standardised formats, or expansion and greater use of existing repositories, such as QsarDB [102], which stores (Q)SAR / QSPR models, would be highly beneficial to modellers enhancing consistency of approaches and preventing duplication of effort.

Knowledge of ADME properties is crucial in the safety assessment of chemicals. Read-across is currently one of the most commonly used alternative approaches for filling data gaps. The read-across assessment framework (RAAF) [103] makes specific reference to the importance of considering the influence of toxicokinetic properties. Hence, having easily accessible

ADME datasets (ideally stored in a centralised repository) would help to elucidate the influence of kinetics and exposure that underlie toxicity. Recently published Integrated Testing Strategies (ITS) and Intelligent Approaches to Testing and Assessment (IATA) have similarly emphasised the importance of ADME information when predicting *in vivo* effects [104]. ADME properties play an essential role in modulating the activity of xenobiotics *in vivo*, governing the temporal concentration at organs of interest which in turn determines the activity / toxicity profile of the compound. As the significance of ADME in determining activity is now widely acknowledged, reliable datasets of these properties, from which robust predictive models can be built, have never been more vital.

Acknowledgment

The founding of the European Chemical Agency (ECHA) Service Contract N.o. FWC ECHA/2013/109 is gratefully acknowledged.

Declaration of interest

All authors state no conflict of interest.

The authors' affiliations are as shown on the cover page. The authors have sole responsibility for the writing and content of the paper.

Bibliography

1. Madden J. Toxicokinetic considerations in predicting toxicity. In: Cronin MTD, Madden JC, editors. In silico toxicology: principles and applications. RSC; Cambridge, UK, 2010:531-57.

*This reference gives an overview on key ADME properties.

2. Hou T, Wang J. Structure-ADME relationship: still a long way to go? Expert Opin Drug Metab Toxicol. 2008;4:759–70.

3. Creton S, Billington R, Davies W, et al. Application of toxicokinetics to improve chemical risk assessment: implications for the use of animals. Reg Toxicol Pharmacol. 2009;55:291-99.

4. EFSA Panel on Food Additives and Nutrient Sources added to Food (ANS); Guidance for submission for food additive evaluations. EFSA Journal. 2012;10(7):2760, doi:10.2903/j.efsa.2012.2760. Available online: www.efsa.europa.eu/efsajournal.

5. Schultz TW, Amcoff P, Berggren E, et al. A strategy for structuring and reporting a read-across prediction of toxicity. Reg Toxicol Pharmacol. 2015;72:586-601.

6. Hou TJ, Wang JM, Zhang W, Wang W, Xu X. Recent advances in computational prediction of drug absorption and permeability in drug discovery. Curr Med Chem. 2006;13:2653–67.

7. Kharkar PS. Two-dimensional (2D) in silico models for absorption, distribution, metabolism, excretion and toxicity (ADME/T) in drug discovery. Curr Top Med Chem. 2010;10:116-26.

8. Mostrag-Szlichtyng A, Worth A. Review of QSAR models and software tools for predicting biokinetic properties. JRC Technical Report EUR 24377 EN, Pubilcations Office of the European Union, Luxemburg 2010.

**This reference gives an excellent insight on the QSAR models predicting biokinetic properties.

9 Cheng FX, Li WH, Liu GX, et al. In silico ADMET prediction: recent advances, current challenges and future trends. Curr Top Med Chem. 2023;13:1273–89.

10. Moroy G, Martiny VY, Vayer P, et al. Toward in silico structure-based ADMET prediction in drug discovery. Drug Discov Today. 2012;17:44–55.

11. Wang Y, Xing J, Xu Y, et al. In silico ADME/T modelling for rational drug design. Q Rev Biophys. 2015;48:488-515.

12. Gola J, Obrezanova O, Champness E, et al. ADMET Property Prediction: The State of the Art and Current Challenges. QSAR Comb Sci 2006;25:1172-80.

13. Wang Y, Xiao J, Suzek TO, et al. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res. 2009;37:623-33.

14. Knox C, Law V, Jewison T, et al. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res. 2011, 39:1035-41.

15. Gaulton A, Bellis LJ, Bento AP, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012;40:1100-7.

16. Judson R, Richard A, Dix D, et al. ACToR-Aggregated Computational Toxicology Resource. Toxicol Appl Pharmaco.l 2008;233(1):7-13.

17. Fourches D, Barnes JC, Day NC, et al. Cheminformatics analysis of assertions mined from literature that describe drug-induced liver injury in different species. Chem Res Toxicol. 2010;23:171-83.

18. Davies M, Dedman N, Hersey A, et al. ADME SARfari: comparative genomics of drug metabolizing systems. Bioinformatics. 2015;31(10):1695–97.

19. https://www.ebi.ac.uk/chembl/admesarfari

20. Cao D, Wang J, Zhou R, et al. ADMET evaluation in drug discovery. 11. PharmacoKinetics Knowledge Base (PKKB): a comprehensive database of pharmacokinetic and toxic properties for drugs. J Chem Inf Model. 2012;52(5):1132-37.

*The important article on available datasets for ADME properties.

21. Orchard S, Al-Lazikani B, Bryant S, et al. Minimum Information about a Bioactive Entity (MIABE). Nat Rev Drug Discov. 2011;10:661–69.

22. Fourches D, Muratov E, Tropsha A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. J Chem Inf Model. 2010;50:1189-1204.

*This article shows that the predictability of QSAR models is directly influenced by various dataset characteristics.

23. Kilkenny N, Browne C, Cuthill WJ, et al. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. PLoS Biol. 2010;8(6):e1000412.

24 Madden JC, Hewitt M, Przybylak K, et al. Strategies for the optimisation of in vivo experiments in accordance with the 3Rs philosophy. Reg Toxicol Pharmacol. 2012;63:140–54.

25. Fu X, Wojak A, Neagu D, et al. Data governance in predictive toxicology: A review. J Cheminform. 2011;3:24.

26. Young D, Martin T, Venkatapathy R, et al. Are the chemical structures in your QSAR correct? QSAR Comb Sci. 2010;27:1337-45.

27. Akhondi SA, Kors JA, Muresan S. Consistency of systematic chemical identifiers within and between small-molecule databases. J Cheminform. 2012;4(1):35.

28. Akhondi SA, Muresan S, Williams AJ, et al. Ambiguity of non‐systematic chemical identifiers within and between small‐molecule databases. J Cheminform. 2015;7:54.

29. Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. Mol Inform. 2010;29:476−488.

30. Jowsey IR, Clapp CJ, Safford B, et al. The impact of vehicle on the relative potency of skin-sensitizing chemicals in the local lymph node assay. Cutan Ocular Toxicol. 2008;27:67-75.

31. Mistry P, Neagu D, Trundle PR, et al. Using random forest and decision tree models for a new vehicle prediction approach in computational toxicology. Soft Comput. 2016;20:2967–79.

32. Ghafourian T, Samarasa EG, Brooks JD, et al. Validated models for predicting skin penetration from different vehicles. Eur J Pharm Sci. 2010;41:612–16.

33. Rowe PH. Essential statistics for the pharmaceutical sciences. Chichester, UK: JohnWiley & Sons; 2007:13–20.

34. MacDonald DD, Ingersoll CG, Berger TA. Development and evaluation of consensus-based sediment quality guidelines for freshwater ecosystems. Arch Environ Contam Toxicol. 2000;39:20–31.

35. Malazizi L, Neagu D, Chaudhry Q. A data quality assessment algorithm with applications in predictive toxicology. Proceedings of the International Multiconference on Computer Science and Information Technology (IPCSIT). 2006:131–40.

36. Gleeson MP, Modi S, Bender A, et al. The challenges involved in Modelling Toxicity Data in silico: A Review. Curr Pharm Des. 2012;18(9):1266-91.

37. Steinmetz FP, Enoch SJ, Madden JC, et al. Methods for assigning confidence to toxicity data with multiple values − identifying experimental outliers. Sci Total Environ. 2014;482:358−65.

38. Roncaglioni A, Benfenati E, Boriani E, et al. A protocol to select high quality datasets of ecotoxicity values for pesticides. J Environ Sci Health B. 2004;39(4):641-52.

39. Steinmetz FP, Madden JC, Cronin MTD. Data quality in the human and environmental health sciences: using statistical confidence scoring to improve QSAR/QSPR modeling. J Chem Inf Model. 2015;55:1739−46.

40. Madden J. Sources of chemical information, toxicity data and assessment of their quality In Cronin MTD, Madden JC, Enoch SJ, Roberts DW, editors. Chemical Toxicity Prediction: Category Formation and Read-Across. RSC; Cambridge, UK, 2013:98-126.

41. Dearden JC. In silico prediction of ADMET properties: how far have we come? Expert Opin Drug Metab Toxicol. 2007;3(5):635-39.

42. Schneider K, Schwarza M, Burkholderb I, et al. "ToxRTool", a new tool to assess the reliability of toxicological data. Toxicol Lett. 2009;189:138-44.

43. Klimisch, HJ, Andreae M, Tillmann U. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regul Toxicol Pharmacol. 1997;25:1-5.

44. Nendza M, Aldenberg T, Benfenati E, et al. Data quality assessment for in silico methods: a survey of approaches and needs. In: Cronin MTD, Madden JC, editors. In silico toxicology: principles and applications. RSC; Cambridge, UK, 2010:59-118.

45. Przybylak KR, Madden JC, Cronin MTD, et al. Assessing toxicological data quality: basic principles, existing schemes and current limitations. SAR QSAR Environ Res. 2012;23:435-59.

46. Frey JG, Bird CL. Cheminformatics and the Semantic Web: adding value with linked data and enhanced provenance. WIREs Comput Mol Sci. 2013;3:465-81.

47. Raies AB and Bajic VB. In silico toxicology: computational methods for the prediction of chemical toxicity. WIREs Comput Mol Sci. 2016;6:147–72.

48. Gleeson MP, Hersey A, Hannongbua S. In-silico ADME Models: A General Assessment of their Utility in Drug Discovery Applications. Curr Topic Med Chem. 2011;11:358-81.

49. Golbraikh A, Muratov E, Fourches D, et al. Data Set Modelability by QSAR. J Chem Inf Model. 2014;54:1-4.

50. Wang J, Hou T. Advances in computationally modeling human oral bioavailability. Adv Drug Deliver Rev. 2015;86:11-16.

51. Davis AM, Riley RJ. Predictive ADMET studies, the challenges and the opportunities. Curr Opin Chem Biol. 2004;8:378-86.

52. Cronin MTD, Schultz TW. Pitfalls in QSAR. J Mol Struc THEOCHEM 2003;622:39-51.

53. Dalby A, Nourse JG, Hounshell WD, et al. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. J Chem Inf Model. 1992;32(3): 244-55.

54. Available from: http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html

55. Available from: https://iupac.org/who-we-are/divisions/division-details/inchi/

56. O'Boyle NM. Towards a universal SMILES representation - a standard method to generate canonical SMILES based on the InChI. J Cheminf. 2012;4:22.

57. Guasch L, Peach ML, Nicklaus MC. Tautomerism of warfarin: combined chemoinformatics, quantum chemical, and NMR investigation. J Org Chem. 2015;80:9900−09.

58. Guasch L, Yapamudiyansel W, Peach ML, et al. Experimental and chemoinformatics study of tautomerism in a database of commercially available screening samples. J Chem Inf Model. 2016;56:2149−61.

59. Yu H, Kühne R, Ebert R-U, et al. Comparative Analysis of QSAR models for predicting pKa of organic oxygen acids and nitrogen bases from molecular structure. J Chem Inf Model. 2010;50(11):1949–60.

60. Lee AC, Yu J, Crippen GM. pKa prediction of monoprotic small molecules the SMARTS way. J Chem Inf Model. 2008;48:2042-53.

61. Hansch C, Leo A, Hoekman D. Exploring QSAR: hydrophobic, electronic and steric constants. ACS, Washington, DC, 1995.

62. Hou T, Xia K, Zhang W, et al. ADME evaluation in Drug Discovery. 4. Prediction of aqueous solubility based on atom contribution approach. J Chem Inf Comput Sci. 2004;44:266-75.

63. Wang J, Krudy G, Hou T, et al. Development of reliable aqueous solubility models and their application in drug-like analysis. J Chem Inf Model. 2007;47:1395-1404.

64. Moda TL, Torres LG, Carrara AE, et al. PK/DB: database for pharmacokinetic properties and predictive in silico ADME models. Bioinformatics. 2008;4(19):2270-71.

65. Wen Ng H, Doughty SW, Luo H, et al. Development and validation of decision forest model for estrogen receptor binding prediction of chemicals using large data sets. Chem Res Toxicol. 2015;28:2343-51.

66. Ajiboye AR, Abdullah-Arshah R, Qin H, et al. Evaluating the effect of dataset size on predictive model using supervised learning technique. Int J Softw Eng Comp Sci (IJSECS). 2015;1:75-84.

67. Okumura K, Lee IP, Dixon RL. Permeability of selected drugs and chemicals across the blood-testis barrier of the rat. J Pharmacol Exp Ther. 1975;194:89-95.

68. Lien EJ. Structure, properties and disposition of drugs. Prog Drug Res. 1985;29:67-95.

69. Sakiyama R, Pardridge WM, Musto NA. Influx of testosterone-binding globulin (Tebg) and tebg-bound sex steroid hormones into the rat testis and prostate. J Clin Endocr Metab. 1988;67:98-103.

70. Sedykh A, Fourches D, Duan J, et al. Human Intestinal transporter database: QSAR modeling and virtual profiling of drug uptake, efflux and interactions. Pharm Res. 2013;30:996-1007.

71. Martins IF, Teixeira AL, Pinheiro L, et al. A bayesian approach to in silico blood-brain barrier penetration modeling. J. Chem. Inf. Model. 2012;52:1686−97.

72. Pham-The H, González-Álvarez I, Bermejo M, et al. The use of rule-based and QSPR approaches in ADME profiling: a case study on caco-2 permeability. Mol Inf. 2013;32:459-79.

73. Available from: http://modem.ucsd.edu/adme/databases/databases_bioavailability.htm

74. Available from: http://modem.ucsd.edu/adme/databases/databases_Pgp_inhibitor.htm

75. Vasios G, Kosmidi A, Kalantzi OJ, et al. Simple physicochemical properties related with lipophilicity, polarity, molecular size and ionization status exert significant impact on the transfer of drugs and chemicals into human breast milk. Expert Opin Drug Metab Toxicol. 2016;12(11):1273-78.

76. Acrobat Adobe. Available from: https://acrobat.adobe.com/uk/en/acrobat/acrobat-pro.html

77. Magnusson BM, Anissimov YG, Cross SE, et al. Molecular size as the main determinant of solute maximum flux across the skin. J Invest Dermatol. 2004;122:993-99.

78. Hewitt M, Madden JC, Rowe PH, et al. Structure-based modelling in reproductive toxicology: (Q)SARs for the placental barrier. SAR QSAR Environ Res. 2007;18(1-2:)57-76.

79. Heller SR, McNaught A, Stein S, et al. InChI - the worldwide chemical structure identifier standard. J Cheminform. 2013;5:7.

80. Pence HE, Williams AJ. ChemSpider: an online chemical information resource. J Chem Educ. 2010;87:1123-24.

81. Open Babel. Available from: http://openbabel.org.

82. KNIME. Available from: http://www.knime.org.

83. Newby D, Freitas AA, Ghafourian T. Decision trees to characterise the roles of permeability and solubility on the prediction of oral absorption. Eur J Med Chem. 2015;90:751-65.

84. Avdeef A. Absorption and drug development: solubility, permeability, and charge state. 2nd edition. John Wiley and Sons, Inc., 2012.

85. Khajeh A, Modarress H. Linear and nonlinear quantitative structure-property relationship modelling of skin permeability. SAR QSAR Environ Res. 2014;25(1):35–50.

86. Votano J, Parham M, Hall LM, et al. QSAR modeling of human serum protein binding with several modeling techniques utilizing structure−information representation. J Med Chem. 2006;49:7169-81.

87. Obach RS, Lombardo F, Waters NJ. Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 670 drug compounds. Drug Metab Dispos. 2008;36(7):1385-1405.

88. Paixao P, Aniceto N, Gouveia LF, et al. Tissue-to-blood distribution coefficients in the rat: utility for estimation of the volume of distribution in man. Eur J Pharm Sci. 2013;50(3–4):526-43.

89. Zhang Y, Li CSW, Ye J, et al. Porcine brain microvessel endothelial cells as an in vitro model to predict in vivo blood-brain barrier permeability. Drug Metab Dispos. 2006;34(11):1935-43.

90. Zhao YH, Le J, Abraham MH, et al. Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure-activity relationship (QSAR) with the Abraham descriptors. J Pharm Sci. 2001;90:749-84.

91. Manga N, Duffy JC, Rowe PH, et al. Structure-Based Methods for the Prediction of the Dominant P450 Enzyme in Human Drug Biotransformation: Consideration of CYP3A4, CYP2C9, CYP2D6. SAR QSAR Environ Res. 2005;16(1-2):43-61.

92. Yap CW, Chen YZ. Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. J Chem Inf Model. 2005;45:982-92.

92. Chohan KK, Paine SW, Mistry J, et al. A Rapid Computational Filter for Cytochrome P450 1A2 Inhibition Potential of Compound Libraries. J Med Chem. 2005;48:5154-61.

94. Paixão P, Gouveia LF, Morais JA. Prediction of the in vitro intrinsic clearancedetermined in suspensions of human hepatocytes by using artificial neural networks. Eur JPharm Sci. 2010;39:310-21.

95. Yap CW, Li ZR, Chen YZ. Quantitative structure–pharmacokinetic relationships for drug clearance by using statistical learning methods. J Mol Graph. 2006;24(5):383–95.

96. Kim M, Sedykh A, Chakravarti SK, et al. Critical evaluation of human oral bioavailability for pharmaceutical drugs by using various cheminformatics approaches. Pharm Res. 2014;31:1002-14.

97. Broccatelli F, Carosati E, Neri A, et al. Novel approach for predicting P-glycoprotein (ABCB1) inhibition using Molecular Interaction Fields. J Med Chem. 2011;54:1740–51.

98. Karlgren M, Vildhede A, Norinder U, et al. Classification of inhibitors of hepatic organic anion transporting polypeptides (OATPs): influence of protein expression on drug-drug interactions. J Med Chem. 2012;55:4740–63.

99. Heller SR, McNaught A, Pletnev I, et al. InChI, the IUPAC International Chemical Identifier. J Cheminform. 2015;7:23.

100. Warr WA. Representation of chemical structures. Wiley Interdiscip Rev Comput Mol Sci. 2011;1:557–79.

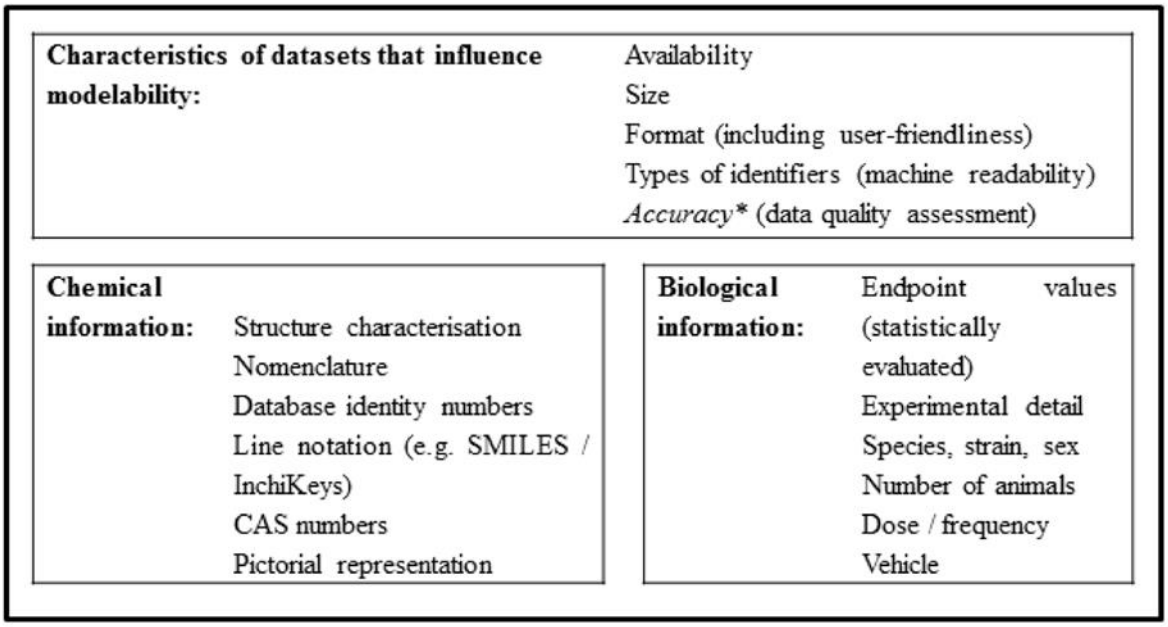
101. Cronin MTD. Finding the data to develop and evaluate (Q)SARs and populate categories for toxicity prediction. In: Cronin MTD, Madden JC, eds. In Silico Toxicology: Principles and Applications. Cambridge, UK: The Royal Society of Chemistry; 2010, 31–58.

102. Available from: http://www.qsardb.org/.

103. European Chemicals Agency (ECHA). 2015. Read-Across Assessment Framework (RAAF), ECHA-15-R-07-EN.

104. Patlewicz G, Kuseva C, Kesova A, et al. Towards AOP Application – Implementation of an Integrated Approach to Testing and Assessment (IATA) into a Pipeline Tool for Skin Sensitisation. Regul Toxicol Pharmacol. 2014;69:529-45.

Figure 1. Factors affecting dataset modelability and the associated chemical and biological information



**NB data quality assessment is not explicitly reviewed here, the reader is referred to the extensive reviews available elsewhere [40, 42-45].*

Table 1. A summary of the ADME properties considered in this investigation.

ADME property	Definition	Property group
ABSORPTION		
% Abs	Percentage of available compound absorbed across (intestinal) barrier	Absorption
% HIA	Percentage that is absorbed across the human gastrointestinal tract	
***In Vitro* (Oral) Absorption Data**		
Caco2	Artificial membrane for predicting absorption - includes paracellular route and active uptake /efflux	
PAMPA	Parallel artificial membrane permeability assay for predicting absorption - models passive diffusion only	
MDCK	Madin-Darby canine kidney epithelial cells used to model absorption	
Skin Absorption		
Kp	Skin permeability coefficient	
Jmax	Maximum rate of flux across / into skin	
Uptake across respiratory tract		
Amount absorbed	e.g. Weight of compound absorbed / kg body weight	
DISTRIBUTION		
%PPB	Percentage of compound bound to plasma proteins	%PPB, fb and fu in plasma
fb	Fraction bound to plasma proteins	
fu	Fraction unbound in plasma (i.e. free fraction)	
fbt	Fraction bound in tissues	fbt, fut in tissues
fut	Fraction unbound in tissues	
Vd	Apparent volume of distribution (Vd) = a hypothetical volume into which a compound distributes; Vdu = the volume for the unbound fraction	Volume of distribution
Vdu		
Ktb	Tissue: blood partition coefficient	
BBB partitioning	Blood brain barrier partitioning, ratio of concentrations between brain and blood (serum / plasma)	
m:p	The ratio of concentration between breast milk and plasma	
Blood:testes	The ratio of concentrations between blood and testes	
CI	Clearance index for placental transfer of compounds usually expressed as a ratio using	

	antipyrine as a marker	CI and TI for placenta transfer
TI	Transfer index for placental transfer of compounds	
***In Vitro* Blood-Brain Barrier (BBB) Partitioning**		
PAMPA-BBB	PAMPA Assay for BBB Partitioning	
BBMEC	Bovine Brain Microvessel Endothelial Cells	*In vitro* BBB
METABOLISM		
Predominant enzyme responsible	Identification of key enzymes (e.g. CYPP450 3A4, 2C9, 2D6 etc.)	
% metabolised	Total percentage metabolised	
% excreted	The percentage of compound excreted unchanged in urine.	Metabolism and % excreted unchanged
Cl / Clh	Cl = volume of blood from which compound is removed in a given time; Clh = Clearance by hepatic route (i.e. metabolism)	
Cyp Inhibition	Inhibition of Cytochrome P450 enzymes	
***In Vitro* Metabolism Data**		
Km	Binding affinities for metabolic enzymes	
Vmax	Maximum velocity of metabolic reaction	*In vitro* metabolism
Cliv	Clearance rate in liver (obtained e.g. using liver slices, microsomes, S9 fraction or liver homogenate)	
ELIMINATION		
Clr	Clearance by renal route (i.e. urinary excretion)	
Cltot	Sum clearance by all routes	Clearance
Composite Parameters		
F	Bioavailability; fraction of dose that enters the systemic circulation	
AUC	Area under concentration time curve	
Cmax	Maximum concentration in blood / plasma	Kinetic data
Tmax	Time to reach max concentration	
$t_{½}$	Half–life i.e. the time taken for the concentration of a compound in the body to fall by half	
Transporter interactions		
Km	Substrate binding affinities (PgP, OATP etc.)	
Ki	Inhibitors (PgP, OATP etc.)	Transporters

Table 2. The summary of assessment of selected ADME datasets.

ID	ADME Endpoints	Availability (Reference)	Size (no. chemicals)	Format	Chemical identifiers
BM 1	Absorption	[83]	932	Excel	Name, CAS, SMILES
BM 2	Caco-2	[72]	1301	Excel	Name, SMILES
BM 3	PAMPA	[84]	290	Pdf	Name
BM 4	MDCK	[83]	246	Excel	Name, CAS, SMILES
BM 5	Kp	[85]	283	Excel	Name
BM 6	Jmax	[77]	278	Gif	Name
BM 7	%PPB, fb, fu in plasma	[86]	1008 (808 +200)	Excel + Pdf	Name
BM 8	%PPB, fb, fu in plasma	[87]	554	Excel	Name, CAS
BM 9	Vd	[87]	670	Excel	Name, CAS
BM 10	Ktb	[88]	143 *in vitro* +196 *in vivo*	Pdf	Name
BM 11	BBB partitioning	[71]	2053	Text	Name, SMILES
BM 12	m:p	[75]	375	Word	Name
BM 13	Blood:testis	[67]	10	Word	Name
BM 14	Blood:testis	[68]	7	Pdf	Name
BM 15	Blood:testis	[69]	6	Pdf	Name
BM 16	CI and TI for placenta transfer	[78]	78 CI + 56 TI	Pdf	Name, CAS
BM 17	In Vitro BBB	[89]	16	Pdf	Name
BM 18	Metabolism and % excretion	[90]	241 (excretion)	Pdf	Name
BM 19	Metabolism and % excretion	[91]	147 (CYP metabolism)	Pdf	Name, SMILES

BM 20	CYP inhibition	[92]	702 (3A4) + 702 (2D6) + 702 (2C9)	Excel	Name
BM 21	CYP inhibition	[93]	87 (1A2)	Pdf	Name
BM 22	*In vitro* metabolism	[94]	94	Pdf	Name
BM 23	Clearance by all routes	[87]	670	Excel	Name, CAS
BM 24	Clearance by all routes	[95]	503	Excel	Name, SMILES
BM 25	Bioavailability F	[73]	1014	Sdf	Name, SMILES
BM 26	Bioavailability F	[96]	995	Excel	Name, SMILES
BM 27	t ½	[87]	670	Excel	Name, CAS
BM 28	Transporter interactions	[70]	3763	Excel	Name, SMILES, CAS
BM 29	Transporter interactions	[74]	1302	Sdf	Name, SMILES
BM 30	Transporter interactions	[97]	1275	Excel	Name, SMILES
BM 31	Transporter interactions	[98]	225	Excel	Name, SMILES

BM - Benchmark
PAMPA - Parallel artificial membrane permeability
MDCK - Madin-Darby canine kidney
Kp - Skin permeability coefficient
Jmax - Maximum rate of flux
PPB - Plasma protein binding
fb - Fraction bound
fu - Fraction unbound
Vd - Volume of distribution
Ktb - Tissue: blood partition coefficient
BBB - Blood brain barrier
m:p – milk:plasma
CI - Clearance index
TI - Transfer index
CYP - Cytochromes P450

t ½ - Half–life
CAS - Chemical Abstracts Service
SMILES - Simplified Molecular Input Line Entry System

Table 3. Characteristics of ideal datasets and their constitutive chemical and biological information

Dataset	**Chemical Information**	**Biological Information**
Freely, publicly-available Large number of compounds covering diverse chemical space Flexible format: information readily downloadable; allows for data manipulation \ transformation; interchangeable with other formats; user-friendly interface Standard identifiers used; consistency between identifiers confirmed Availability of the reference to the original study Accurate: quality controlled / quality assured data	Structures appropriately characterised and "cleaned" as necessary to ensure suitability for modelling e.g. removal of metallics, salt forms, tautomers etc. Identifiers to include: Nomenclature SMILES string InChIKey CAS number Pictorial representation (in a chemical standard format)	Endpoint values stated with consistent units; statistical validity of measured values confirmed Adequate experimental detail reported Generic and endpoint-specific sources of experimental variability identified and minimised: e. g. species, sex, strain, number of animals used, dosing regimen. Vehicle reported (same vehicle used where possible)

SMILES – Simplified Molecular Input Line Entry System
InChIKey - International Chemical Identifier Key
CAS - Chemical Abstracts Service