Orchard, S., Al-Lazikani, B., Bryant, S., Clark, D., Calder, E., Dix, I., Engkvist, O., Forster, M., Gaulton, A., Gilson, M., Glen, R., Grigorov, M., Hammond-Kosack, K. E., Harland, L., Hopkins, A., Larminie, C., Lynch, N., Mann, R. K., Murray-Rust, P., Lo Piparo, E., Southan, C., Steinbeck, C., Wishart, D., Hermjakob, H., Overington, J. and Thornton, J. 2011. Minimum information about a bioactive entity (MIABE). *Nature Reviews Drug Discovery.* 10, pp. 661-669.

OPINION

Minimum information about a bioactive entity (MIABE)

Sandra Orchard, Bissan Al-Lazikani, Steve Bryant, Dominic Clark, Elizabeth Calder, Ian Dix, Ola Engkvist, Mark Forster, Anna Gaulton, Michael Gilson, Robert Glen, Martin Grigorov, Kim Hammond-Kosack, Lee Harland, Andrew Hopkins, Christopher Larminie, Nick Lynch, Romeena K. Mann, Peter Murray-Rust, Elena Lo Piparo, Christopher Southan, Christoph Steinbeck, David Wishart, Henning Hermjakob, John Overington and Janet Thornton

Abstract | Bioactive molecules such as drugs, pesticides and food additives are produced in large numbers by many commercial and academic groups around the world. Enormous quantities of data are generated on the biological properties and quality of these molecules. Access to such data — both on licensed and commercially available compounds, and also on those that fail during development — is crucial for understanding how improved molecules could be developed. For example, computational analysis of aggregated data on molecules that are investigated in drug discovery programmes has led to a greater understanding of the properties of successful drugs. However, the information required to perform these analyses is rarely published, and when it is made available it is often missing crucial data or is in a format that is inappropriate for efficient data-mining. Here, we propose a solution: the definition of reporting guidelines for bioactive entities — the Minimum Information About a Bioactive Entity (MIABE) — which has been developed by representatives of pharmaceutical companies, data resource providers and academic groups.

The process of identifying and developing molecules with useful bioactive properties, such as pharmaceutical agents, pesticides and food additives, is fraught with difficulty. In the pharmaceutical industry only a very small percentage of investigational new drugs will make it through to approved clinical usage, for reasons that include lack of efficacy, unexpected side effects and toxicity, and undesirable drug–drug interactions. Active pesticides that have undesirable side effects against organisms other than their original target will also not make it to the market. Food additives such as plant sterols are subject to strict regulations on their safety and possible off-target effects.

Published reports on the properties of 'failed' molecules, in addition to detailed information on those that become fully licensed and commercially available, are crucial for developing an understanding of how improved molecules may be developed. Details of the molecular structure and mechanism of action of molecules may give clues as to how related analogues could be developed that modulate the same target, but with improved effectiveness or increased specificity. A full disclosure of observed toxicity or an understanding of the pharmacokinetic properties of an agent may help to improve these properties in subsequent generations of molecules. Even those molecules that fail at an early stage in the development process may be useful as tools; for example, in aiding the validation of potential new targets that are identified through genomic and/or proteomic studies of disease.

Another area in which integrating knowledge on bioactive entities could be valuable is in identifying proteins that could be amenable to modulation by small molecules. In 2002, Hopkins and Groom[1] introduced the concept of the 'druggable genome' to describe the portion of the human genome that encoded proteins that were both linked to disease and were tractable targets for small-molecule drugs. Since then, progress has been made with computational approaches to calculate the druggability of proteins and to predict druggable proteins[2–5]. In 2006, Overington *et al.*[6] estimated that the number of protein targets for approved drugs at the time was as low as 266 out of the 20,400 protein-coding genes[7] in the human genome (a further 58 targets were either from pathogenic organisms or they were non-protein molecules). Similar models may be developed to assess the susceptibility of proteins encoded by plant or microbial genomes to modulation by small molecules. Mining computationally accessible resources of data on bioactive entities could provide synergistic benefits for research and development across the fields of drug discovery, pesticide science, cosmetics, nutraceuticals, metabolomics and toxicology.

At present, however, much of the substantial volume of data that would aid predictions of protein targets that could be successfully modulated, and the development of new or improved molecules to modulate these targets, resides in proprietary databases. This is especially the case for agents that have subsequently failed at some stage of the discovery process (for example, owing to poor pharmacokinetics or toxicity). Nevertheless, owing to the current productivity crisis and decreasing budgets for early-stage research, the pharmaceutical, biotechnology and food industries are increasingly considering the precompetitive release of compound-related bioactivity data into public repositories, recognizing that this could not only provide a general benefit for research but also a commercial advantage as data only need to be collected once. The manual curation of data is an expensive process and requires resources that are often not available in even the largest pharmaceutical

companies. In addition, a single company has only a limited range of in-house products that target an equally restricted number of targets. By integrating knowledge from other commercial and academic groups, a broader understanding of the viability of a potential target protein may be reached before expensive resources are committed to a research programme.

This precompetitive activity can also have wider benefits; for example, in academic or non-profit discovery and in the development of drugs for orphan and tropical diseases. A recent encouraging example has been the release of data by GlaxoSmithKline, Novartis and St Jude Children's Research Hospital, USA, on over 14,000 compounds that are known to be active against the parasite that is responsible for transmitting malaria; these data are now available in the ChEMBL database. The same database was used in a chemogenomic screen that identified *Schistosoma mansoni* proteins against which existing drugs may be active, thus providing hope that known therapeutics could be used to treat the neglected tropical disease schistosomiasis, which affects 210 million people in 76 countries[8].

However, to fully understand the growing wealth of data on bioactive compounds, the context, methods, results, conclusions and detailed background information pertaining to experiments also need to be included. The diversity of experimental designs and analytical techniques is increasing as new methods are being developed and the scale of data production is rising owing to the widespread application of automation. This makes the formal specification of these metadata — 'data about the data' — of increasing importance. By being associated with the results, these metadata make both the biological and methodological contexts of the experiment explicit. The archetype of such a specification is the 'Minimum Information about a Microarray Experiment' (MIAME)[9]. Many journals and funding agencies now require authors reporting on microarray-based transcriptomics experiments to comply with the MIAME checklist as a prerequisite for publication. The adoption and development of such specifications has had a considerable impact beyond simply increasing the comprehension and comparability of journal articles, the most important aspect of which is facilitating the transfer of data from journal articles into databases (that is, converting unstructured data into structured data) in a form that enables data mining across combined data sets.

We therefore propose a new document — the 'Minimum Information About a Bioactive Entity' (MIABE) — that is predominantly concerned with, but not restricted to, bioactive chemical compounds. We believe the timing of this is apposite for several reasons. The first reason is the ongoing revolution in databases that hold information on the bioactivity of small molecules. In particular, the embedding of the 'missing entity' of chemical structures in databases such as the Chemical Entities of Biological Interest (ChEBI) database[10] and the PubChem project[11] during the past decade has made it possible to search across biological effects, protein names, sequence data and chemical information. New technologies have emerged that can be used to directly interconnect the data sets in different types of databases, thereby improving the power of any search. Second, results from high-throughput screening, as well as other types of bioactivity screening data, that are directly linked to information on chemical structures are increasingly being deposited in public repositories such as the PubChem Bioassay[11], ChemBank[12] and ChEMBL databases. Third, drug discovery and agricultural studies are no longer limited to the commercial sector. Publicly funded initiatives such as the US National Institutes of Health's Molecular Libraries Program, the Innovative Medicines Initiative and the European Union's EU-Openscreen project aim to both enhance chemical biology efforts through high-throughput screening and speed up the development of better and safer medicines. The resulting data from these programmes belong in the public domain and it is in everyone's interest to ensure that these data are made available in the most appropriate format and with the correct metadata available. Last, increasing legislation is requiring that information on bioactive compounds should be more readily available. For example, on 1 June 2007 the REACH (Registration, Evaluation, Authorisation and Restriction of Chemical substances) legislation came into force across the European Union. This legislation requires that additional information on chemicals should be made available, depending on the quantity imported or manufactured, and it has been estimated that 30,000 chemicals may need to be re-evaluated in accordance with these requirements[13].

It is hoped that the publication of these proposed standards will encourage the deposition of data that fulfil the MIABE criteria into public databases such as ChEMBL or PubChem, not only by the academic community but increasingly also by companies. Following on from the period during which these guidelines were publicly available for feedback (see below), we also hope that highlighting their availability through publication will now provide the opportunity for broad community feedback and adaptation if needed. Once a stable version of these guidelines has been established, however, it is intended to make updates as infrequent as advances in technology will allow, thus providing long periods of stability to encourage the adoption and implementation of the guidelines.

MIABE: principles and process

In order to derive the maximum benefit from the publication of data on bioactive entities, it is important that key information is included in each paper such that the properties of these molecules are fully represented, their effects on biological systems (both positive and negative) are accurately detailed and factors that may contribute to the activity of the molecule are stated.

To this end, we brought together representatives of life science companies, data resource providers and academic groups to develop a checklist of the information that is considered to be important to include with published data sets on bioactive entities; this is the minimum information that is required for the activity of a molecule to be both fully understood and compared with other molecules that either share a common chemical structure or a similar mechanism of action. Initial versions of the MIABE standard were produced through the Pharmaceutical Industry Forum hosted at the European Bioinformatics Institute, and were made available for community input through pre-publication on the Human Proteomics Organization Proteomics Standards Initiative (HUPO-PSI) website[14] and also via the Minimum Information for Biological and Biomedical Investigations (MIBBI) portal (see the MIBBI portal website). Extensive consultation was then undertaken until the current version of the guidelines was deemed to be appropriate for publication.

It is intended that the MIABE guidelines will be adhered to by anyone planning on publishing papers reporting on bioactive entities, by the implementers of resources such as databases that hold this information and by any body or institution that funds such discovery work and requires the results to be published at the end of the grant period. As with other similar reporting guidelines[9], the MIABE checklist adheres to

the criteria of sufficiency (a reader should be able to understand and critically evaluate the interpretation and conclusions, and support their experimental corroboration) and practicability (the guidelines should not be so burdensome as to prohibit their widespread use).

It should be noted that the scope of this document is limited to data on bioactive entities that would be regarded either as preclinical (for potential drugs) or in early-stage development (for other types of bioactive molecules); a discussion of the publication of clinical data is a subject that is more appropriate for a separate effort. It is also recognized that the full list of requirements described in the document covers the entire path taken by a molecule from synthesis to preclinical development as a potential drug (or equivalent). Such data may not be available for many compounds that fail to fulfil one or more of the criteria that are necessary for their development into a potential drug, or for compounds that were not generated with such a goal in mind, but all of the data generated on such compounds should appear in any publication on their activity, to maximize the value of these compounds as research tools. Similarly, data on a successful agent may be published in multiple papers. Early-stage investigations — for example, a report on the results of a high-throughput screen — may not have included studies that generate all of the physicochemical data listed in the MIABE guidelines. These guidelines are not intended to provide an additional burden to researchers by requiring the generation of data merely for the sake of publication, but in cases in which data have been generated they should be fully and clearly reported with the appropriate metadata included. Finally, it should also be remembered that although this document focuses on bioactive agents, data on inactive, closely related analogues are often of equal value, and can provide negative controls and information for those trying to build bioactivity into a particular molecular scaffold. Such data should also be fully reported using these guidelines.

Reporting requirements in MIABE

The current MIABE checklist is provided in TABLE 1, and details on two exemplified compounds from a relevant journal article[15] have been presented as a MIABE-compliant document in Supplementary information S1 (table) and Supplementary information S2 (table). Further discussion of the aspects of the checklist is provided in the sections below.

Data formats and related initiatives. It has not previously been standard practice when reporting on bioactive molecules to consider data exchange formats other than the published paper. However, as public data repositories become established, the requirement to exchange data between them — or for users to download non-redundant data sets in a common format — will become increasingly important. In the molecular interaction field, such a format — the HUPO PSI-Molecular Interactions (MI) XML 2.5 interchange format — exists and has been publicly available and widely used for several years. This format can capture extensive details about many types of interactions, including interactions between bioactive entities and their target molecules, the biological role of each molecule within that interaction, a detailed description of interacting domains and the kinetic parameters of the interaction[16]. The ability to describe the structure of individual molecules and to carry out metadata analysis on that molecule is also inherent to this format. This format is supported by data management and analysis tools and has been adopted by major interaction data providers and tool developers, and used in visualization and analytical software. In addition, a simpler tab-delimited format — MITAB2.5 — has been developed for the benefit of users who require only minimal information in an easily accessible configuration.

It has been suggested that supporting the continued development of this format, rather than attempting to 'reinvent the wheel', will be the most practical approach for fulfilling the MIABE goals. This will allow producers of drug–target information to merge their data with existing information in the molecular interaction databases and use HUPO PSI-MI XML 2.5-compliant resources, such as Cytoscape, to visualize small-molecule data in conjunction with cellular interactomes or pathways. For example, by mapping much of its data to the HUPO PSI-MI standards and formats, the ChEMBL database has made this content available as a Proteomics Standard Initiative Common Query Interface (PSICQUIC) web service[17]. This allows users of the Reactome pathways database[18] to query for small molecules that modulate a specific point in a pathway. Supplementary information S2 (table) contains the HUPO PSI-MI XML file for the publication[15] from which the examples in Supplementary information S1 (table) were taken.

Standardization of the reporting and collection of data is now being encouraged across the biomedical field in an effort to improve data quality and availability. To manage this process, the MIBBI project (see the MIBBI website) has been established, which maintains a web-based, freely accessible resource for checklist projects such as MIABE. MIBBI, which is managed by representatives of its participant communities, provides access to checklists (as well as complementary data formats, controlled vocabularies, tools and databases) and ensures that new efforts are not redundant or overlapping with an existing resource[19]. If the MIABE guidelines overlap with existing resources, the work of these resources is referred to, rather than repeated. For example, the MIABE guidelines refer to the Minimum Information required for reporting a Molecular Interaction Experiment (MIMIx) standard[20], the Standards for Reporting Enzymology Data (STRENDA) guidelines, the Minimum Information about a Protein Affinity Reagent (MIAPAR) standard[21] and the Minimum Information about a Cellular Assay (MIACA) standard. There are ongoing efforts at the MIBBI Foundry to provide a roadmap for users of all 'Minimum Information' documents, in which more than one checklist may be relevant to the information described in their publication or database submission, which will enable users to consult a single list of requirements instead of reading multiple documents.

Controlled vocabularies. Where possible, the MIABE guidelines propose using existing controlled vocabularies or ontologies to describe entities, processes and conditions detailed within a paper. This benefits the reader as it reduces the possibility of ambiguity if a term with multiple meanings is used, and also aids the curation of data, as well as subsequent search and analysis. Several such controlled vocabularies are available on the Open Biological and Biomedical Ontologies website. These controlled vocabularies may be used to describe tissues, diseases and molecular interactions (including enzyme–substrate interactions). Shuffenhauer *et al.*[22] carried out an early attempt to develop an ontology that is specific to bioactive small molecules. The existing HUPO PSI-MI controlled vocabulary[14] that is used to annotate the format described above has already been extended to allow for a full description of the properties of a bioactive molecule, and further input into the development of this resource is welcomed. It is recommended that the HUPO PSI-MI

Table 1 | **Minimum information about a bioactive entity**

Information	Notes
Responsible person or role	
Contact person; organization; contact e-mail	The (stable) primary contact person for this data set should be provided; this could be the individual carrying out the experiment, the head of a laboratory or a line manager. In cases in which responsibility rests with an institutional role (for example, one of a number of duty officers) rather than an individual, the official name of the role rather than the name of an individual should be provided. In all cases, the affiliation and stable contact information should be given
Molecule properties	
Primary name (and synonyms)	Authors should select an appropriate name by which the molecule should be known — for example, the international nonproprietary name (INN) or research codes — but also list any synonyms that are already widely used in the public domain, if these are not already listed within a public domain database. If the molecule is described in a public database, the relevant accession number in that resource and the database name should be given. Internal compound designations or research codes need to be checked for homonym clashes, explicitly listed and clearly matched with either the corresponding public database identifier or a standard structural representation such as SMILES or InChI
Molecule type (MI:0313)	Authors should specify the molecular type of the bioactive entity: for example, whether it is a small molecule, a protein, an antibody, and so on. It is recommended that the terms taken from 'interactor type' in the HUPO PSI-MI controlled vocabulary are used (see the Ontology Lookup Service for further information)
Chemical IUPAC name (MI:2007)	The standard chemical name for the drug should be specified
Chemical structure (MI:2009)	An image of the structure or sequence of the molecule should be provided, as applicable. A corresponding accession number from an externally available resource is also acceptable (for example, ChEBI, PubChem or UniProtKB), provided the molecule is an exact match to the bioactive entity being described. If a molecule is a fragment or an engineered product of the entity that is described in the database, changes to the given structure or sequence should be described in full detail. For small molecules, exact details of stereochemistry should be included. If a molecule contains specific isotopes of an atomic element and not the natural isotope distribution, this should be stated
Standard InChI string and key (MI:0970)	The InChI and InChI key should be provided for all small molecules. The InChI string is derived from the structure of a compound.The InChI key is a fixed-length format that is directly derived from InChI. It is based on a strong hash (SHA-256) algorithm of an InChI string and allows for efficient database indexing. The programme and the version of the programme that has been used to generate the InChI key should be stated
Chemical salt	If different salts (for example, chloride or mesylate) of a chemical are prepared, it should be clearly stated throughout which salt has been used in each experiment. Other known forms of a molecule — for example, hydrates — should also be described
Prodrugs	If a bioactive entity has been synthesized as a prodrug, both the original and final forms of the molecule should be named and described as above. The route by which the molecule is metabolized to its final form should also be described, if this is known
Molecule production	
Chemical synthesis and/ or molecule production	The synthetic route by which a molecule, or series of molecules, has been produced should be fully described or referenced, if known. If the molecule is a natural product, the source, extraction method, estimated purity and yield should be stated. If it has been purchased, the manufacturer and product number should be stated
Percentage purity	The percentage purity of any molecule should be given and the method by which this measure has been derived should be briefly described
Physicochemical properties	
Molecular weight (MI:2025)	The molecular weight of the compound (clearly stating whether salt and/or water molecules are included or not) should be provided
Water solubility (experimental) (MI:2027)	The water solubility of the compound should be given (in mg per ml or g per l)
LogP/hydrophobicity (experimental) (MI:2029)	The logarithm of the water/octanol partition coefficient (logP) or hydrophobicity score should be given. For charged molecules, a logD against a defined buffer/organic phase should also be supplied*
In vitro *cell-free assays*‡	
Name of primary target	Protein nomenclature should be taken from an external resource, such as UniProtKB, or it should be widely accepted in the literature. Gene nomenclature should either be the approved nomenclature for that species — for example, from the HGNC — or it should be taken from a public domain resource such as Ensembl. Names must either be accompanied by an accession number from an external resource (for example, Ensembl, GeneID or UniProtKB) or a statement giving the species of origin of the gene and/or protein product. If the assay is directed against a heteromeric protein complex that is known to comprise multiple proteins, the common name of this complex should be used (for example, 20S proteasome or γ-secretase)
Assay details	A free-text description of the assay should be provided, giving enough information for the reader to understand and, if necessary, reproduce the assay conditions. This should include a clear description of the substrate, preferably including an accession number from an external resource such as UniProtKB or ChEBI
Assay parameters	Any parameter of the assay (for example, pH, temperature or time course) that may affect the final result should be described in full detail

Table 1 cont. | **Minimum information about a bioactive entity**

Information	Notes
In vitro *cell-free assays*[‡] *cont.*	
Delivery systems	A description of the solvent that is used to dissolve the test molecule (for example, DMSO), and a statement of the final concentration of that solvent in the assay should be given. Details of control samples should also be included
Results	Where possible, results should be given as a numerical value with experimental error, rather than — or in addition to — a free-text description. If a derived kinetic value (for example, K_i) is given, the method by which it was derived should be stated, along with any range in values observed. Values that are less dependent on assay conditions should be quoted by preference — for example, K_i values should be quoted rather than IC_{50} values
Secondary gene targets	Results obtained against all possible targets should be given, even when they are negative, to give as complete a profile as possible when describing the activity of the bioactive agent
Cellular assays[§]	
Cell type	Details of the cell line or the source of primary cell culture should be given, such that the reader could obtain or prepare the same cell type. If a cell line is derived in-house, details should be given and a statement made if the cell line is available on request. Any change to the wild-type state of the cell — for example, the transfection of additional or mutated genes — should be noted
Culture conditions	A full statement should be given of the culture conditions, including media, passage number and the perceived state of the cell at the time of the experiment
Agonists or antagonists	A full description should be given of all agonists (or antagonists) added, including the concentration used, length of exposure of the cell to the agonist or antagonist and details of any delivery vehicle used
Results	Results should be described in full detail and, where possible, accompanied by numerical values assessing the activity of the agent and experimental error. If a derived value is given, the method by which it was derived should be stated, along with any range in values observed. If a statistical assessment of the data is made, details of the method used should also be given
Secondary cellular assays	Results obtained against all possible targets should be provided, even when they are negative, to give as complete a profile as possible when describing the activity of the bioactive agent
Toxicological observations	If an effect of potential toxicological relevance is observed — for example, cell death — the effect and the concentration of the agent above which this effect was observed should be recorded
Whole-organism studies[‖]	
Animal studies	For all animal studies the species, strain, sex, age and weight of the animals should be stated. Any feeding or housing conditions that deviate from the norm should be recorded, as should any procedure that may have caused stress to the animal during the course of the experiment. If an animal is anaesthetized or treated with any other agent in addition to the bioactive entity under investigation, full details of this agent should be recorded
Plant studies	For all plant studies the species, strain and life-cycle stage of the plant should be stated. Any nutrient or housing conditions that deviate from the norm should be recorded, as should any procedure that may have caused stress to the plant during the course of the experiment. If a plant is treated with any other agent in addition to the bioactive entity under investigation, full details of this agent should be recorded
Fungal studies	For all fungal studies the species, strain and life-cycle stage of the fungus should be stated. Any nutrient or housing conditions that deviate from the norm should be recorded, as should any procedure that may have caused stress to the fungus during the course of the experiment. If a fungus is treated with any other agent in addition to the bioactive entity under investigation, full details of this agent should be recorded
Disease models	A full description should be given of how the disease condition was induced, including timings, drug concentrations and vehicles used, if appropriate. The treatment of control groups should also be described in full detail. If the condition that is induced is a model (or a partial model) of a disease, it should be made clear which disease (or group of diseases) this model is believed to represent
Dosing route	The route by which a drug is delivered (for example, oral gavage) should be clearly stated
Dosing schedule	Full details of any dosing schedule, including vehicle and timings, should be given. Any period of respite from the drug should also be recorded
Results	Results should be described in full detail and, where possible, accompanied by numerical values assessing the activity of the agent and experimental error. If a statistical assessment is made of the data, details of the method used should also be given. If a series of parameters are used as the end point of the disease model, the effect of the agent on each parameter should be described in full detail and the statistical significance of the effect on each parameter should be separately calculated
Toxicological observations	If a potential toxicological observation is made (pre- or post-mortem), the effect and the concentration of the agent above which this effect was observed should be recorded, together with experimental error. General indicators of animal welfare, such as changes in body weight, should be also recorded. Any change in a physiological parameter that is observed during a study should be noted
Drug–drug interactions	All observed drug–drug interactions should be noted. The method by which these interactions were identified and the dosing regime and vehicle that was used for each agent should be stated

Table 1 cont. | **Minimum information about a bioactive entity**

Information	Notes
Pharmocokinetic studies¶	
Absorption	The methods by which drug absorption was measured must be fully described, as detailed above. This description should include predictive systems, such as Caco-2 cell line data, in addition to actual measurements of the compound in animal studies. Details of any additional experimental factors that may affect the rate of absorption — for example, when the drug was delivered with respect to feeding times — should also be recorded
Protein binding	Estimates or measures of protein binding should be given, with appropriate experimental methodology, as detailed above. If an isolated protein has been used — for example, serum albumin — either an accession number or the species of origin of the protein should be stated. If variant data are included, variants should be accurately mapped to an underlying protein sequence (of a given accession number) and crossreferenced to the dbSNP, if possible
Dosing route	The route by which a drug is delivered (for example, oral gavage) should be clearly stated
Dosing schedule	Full details of any dosing schedule, including vehicle and timings, should be given. Any period of respite from the drug should also be recorded
Half-life	The half-life of a drug following both single and multiple dosing should be stated, and the method by which this was measured and any factors that may affect this value should be recorded
V_{max}	The V_{max} of a drug following both single and multiple dosing should be stated, and the method by which this was measured and any factors that may affect this value should be recorded
Volume of distribution	The volume of distribution should be calculated and quoted in litres
Bioavailability	For drugs that are delivered by non-intravenous means, the percentage bioavailability should be given
Metabolism	The methods by which drug metabolism was measured must be fully described, as detailed above. This description should include predictive systems, such as cytochrome P450-isolated enzyme data, in addition to measurements of the actual compound in animal studies. If an isolated protein has been used — for example, a cytochrome P450 enzyme — either an accession number or the species of origin of the protein should be stated. If variant data are included, variants should be accurately mapped to an underlying protein sequence (of a given accession number) and crossreferenced to the dbSNP, if possible. If a compound proves to be inhibitory to a metabolizing enzyme such as a cytochrome P450 enzyme, these data should also be included. Details of any additional experimental factors that may affect the rate of metabolism — for example, when the drug was delivered with respect to feeding times — should also be recorded
Metabolites	All observed drug metabolites should be identified, including intermediates that are further metabolized prior to excretion. The method by which each metabolite was identified, its concentrations and any observed changes in metabolite composition following repeated dosing should be observed
Excretion	The excretion route of the parent compound and any identified metabolite should be identified. The rate of compound (and metabolite) excretion should also be given

*LogP (partition) or LogD (distribution coefficient) are the logarithms of the ratio of concentrations of a compound in the two phases of a mixture of two immiscible solvents at equilibrium. ‡Defined as the testing of the activity of the agent against isolated or partially purified biomolecules. §Defined as the testing of the activity of the agent in whole-cell systems. ||Defined as the testing of the activity of the agent in whole-organism systems. ¶Defined as a determination of the fate of substances that are administered externally to a living organism. ChEBI, Chemical Entities of Biological Interest; dbSNP, Single Nucleotide Polymorphism Database; DMSO, dimethyl sulphoxide; HGNC, HUGO Gene Nomenclature Committee; HUPO-PSI-MI, Human Proteomics Organization Proteomics Standards Initiative Molecular Interactions; IC_{50}, half-maximal inhibitory concentration; InChI, IUPAC International Chemical Identifier; IUPAC, International Union of Pure and Applied Chemistry; K_i, inhibition constant; MI, molecular interaction; SMILES, Simplified Molecular Input Line Entry Specification; UniProtKB, UniProt Knowledgebase; V_{max}, velocity of enzyme-catalysed reaction at infinite concentration of substrate.

controlled vocabulary terms be used as the primary source of controlled vocabulary terms for annotating bioactive entity data, with additional terms sourced from additional vocabularies when required. The HUPO PSI-MI XML file in Supplementary information S2 (table) is fully annotated using such terms.

An ontology for the description of drug discovery investigations has also been published[23]. Researchers may also wish to incorporate terms from this controlled vocabulary, which allows a clear definition of the various stages of compound progression into database depositions or descriptions of their work.

Molecule structure and nomenclature. The representation of bioactive entities in the literature varies in both quality and depth. In many publications there is an image-based representation of the core chemical structure, with substituents listed in a separate table and often accompanied by preliminary biological assay data. Journals tend to recommend or insist upon a particular nomenclature system, but do not enforce this at the level of the individual molecule, which is often only described using an originator-specific identifier. This makes data almost impossible to automatically mine, and currently this information can only be harvested for computational searching by manual curation.

Several algorithms for information-reduced linear notation for chemical structures have been developed, including SMILES (Simplified Molecular Input Line Entry Specification)[24] and International Union of Pure and Applied Chemistry (IUPAC) International Chemical Identifier (InChI) strings or keys[25] (see the IUPAC website), which are computationally accessible but not appropriate for use in free text. It is recommended that the 'standard' InChI key and strings be used, as these remove ambiguity that could arise from user-settable options for tautomerism and stereochemistry. These are supported by the InchI Trust as being the most suitable algorithms for use in search engines and software applications. The InChI key for molecules that have been studied for a research paper should be supplied as part of the publication, either in the main text, supplementary material or in a parallel deposition in a public domain database.

If the molecule is synthesized as a prodrug, both the initial structure of the molecule and

the final bioactive entity should be described. If it is known, the route by which the molecule is metabolized to its final form should be described. If the entity is known to have metabolites that themselves show *in vivo* activity, these should also be detailed. Drug activities, even at the research phase, are often influenced by the choice of counter-ion (salt) or vehicle (such as vegetable oils). As such, the testing results should include these data where possible, particularly in situations in which pharmacokinetics (for example, in intact animal studies) and/or absorption (for example, in cell-based assays) could influence drug concentration at the target.

As a drug progresses to the market, it is not uncommon for several different salts to be created and used in different assays, or for different formulations to be developed for marketing. To ensure that there is no ambiguity in the literature, the form of the molecule being described in the publication should be clearly indicated and the nomenclature used should be as informative and unambiguous as feasibly possible. If it is known or suspected that the molecule has solid-state properties that affect its behaviour — for example, crystal polymorphs — these should be described, if possible. In the case of known molecular structure heterogeneity, a representative molecular structure or structures should be given, with an accompanying explanation of the possible ambiguity.

Increasingly, biomolecules such as recombinant peptides and proteins are being used in medicine. Where possible, these molecules should be referenced to an external resource such as the UniProt Knowledgebase[7] for protein sequences, but additional data such as the mapping of peptide ligands to large precursors should also be provided. Similarly, the presence of tags, fusion proteins, post-translational modifications such as glycosylation or disulphide links, and synthetic modifications such as conjugation to polyethylene glycol (PEG) should be described. If the biological activity depends on a particular engineered form of the original sequence (for example, the replacement of one or more amino acids), this should be clearly indicated. If the protein in question is an antibody, the reader should be referred to the MIAPAR guidelines[21] for a detailed description of required metadata for such molecules.

Molecule synthesis, isolation or purchase. Another aim of the MIABE suite of documents is to enable studies to confirm or refute a given result. In the case of chemically synthesized molecules, the synthetic route should be given in enough detail for a laboratory to have a reasonable chance of reproducing the process described and producing the range of compounds listed in the article. Similarly, if the molecule has been isolated from a particular organism or isolate, the source, growth conditions and purification procedure should be detailed or referenced if known. For all compounds, this should be accompanied by an indication of purity and identity, and the methods by which this has been ascertained; for example, nuclear magnetic resonance or mass spectrometry.

Purchased compounds should be described by giving the manufacturer's name and the appropriate catalogue and batch numbers. Should the molecule already be in common use and fully described by a public resource such as ChEBI[10] or PubChem[11], an accession number will be sufficient to refer back to much of these data but any additional information on subsequent confirmation of identity should be included.

Physicochemical properties. Physicochemical parameters can be qualitative and/or quantitative; if they are quantitative, units should be given. In both cases, the experimental conditions under which the data were generated (for example, temperature and pH) should be stated.

Only experimental parameters are included in the minimum requirements; however, if calculated parameters are given in a publication, the method by which they are calculated should be described.

In vitro *assays.* The MIABE guidelines describe all *in vitro* assays, including mechanism-of-action studies, studies to ascertain off-target activities and drug metabolism studies, as well as cell-free and cellular assays. More detailed guidelines exist in particular areas — such as the STRENDA guidelines for isolated enzyme studies and the MIMIx guidelines for interaction studies[20] — and these should be consulted with the MIABE documentation. Common to all of these documents, however, is a clear requirement that both the compound and the target should be clearly and unambiguously identified (for example, by a sequence database accession number), and that the assay method should be described in enough detail for researchers to both fully comprehend and, if appropriate, reproduce the conditions locally. If a series of compounds has been assessed for a structure–activity relationship, each experimental value should be clearly matched to the structure of the specific compound for which it has been generated. In the case of library screening, Inglese *et al.*[26] have published additional guidelines that are complementary to the MIABE guidelines.

For cell-based assays, a more detailed list of requirements has been drawn up by the MIACA group (see the MIACA homepage), which should also be consulted. It is important that all changes to the wild-type cellular proteome, such as the transfection of genes coding for mutated or tagged proteins, are clearly documented as changes in protein expression patterns may affect both the cellular phenotype and the response to bioactive entities. In particular, if a protein is overexpressed in a recombinant cellular assay and is the primary focus of that assay, the target (or targets) should be clearly and unambiguously identified, as for *in vitro* cell-free assays. For cellular assays in which the readout is a biological process for which the molecular pathway that is the hypothesized focus of the assay is well understood, the biological process, pathway and primary protein target (or targets) should be stated using existing controlled vocabularies such as the Gene Ontology[27].

In vivo *assays.* In cases in which models are intended to represent a disease condition or a process that is common to multiple diseases — such as inflammation — this should be described as fully as possible, potentially by reference to a disease ontology, such as the one available at the Disease Ontology Community Wiki website. If the assay uses a model organism that has undergone genetic manipulation, such as gene knockout or transgene insertion, and the behaviour of the bioactive entity is compared with reference to the wild-type strain, the potential target (or targets) in the genetically modified organism should be clearly identified. All observations should be fully reported, including pre- and post-mortem and subsequent analyses. In particular, data relating to potential off-target and toxicological effects are commonly not reported, but hold valuable information that may aid the design of subsequent generations of bioactive molecules.

In vivo assays also have ethical considerations. It is not the purpose of these guidelines to discuss these ethical considerations, which are often specific to the country of the originating research facility. Publication requirements with regard to ethical considerations are also provided by journals.

Box 1 | **Frequently asked questions about MIABE**

Below we clarify common questions about using and contributing to the Minimum Information about a Bioactive Entity (MIABE) guidelines.

What is MIABE and what is it for?

- MIABE is a formal list of the items of information that should be provided when describing the synthesis and subsequent analysis of any potentially bioactive entity (a molecule that is designed to show activity in a biological assay)
- MIABE is a checklist of all the information that should support the description of each compound in either a journal article or database submission. Although it is acknowledged that some data may only be available though reference to an earlier article, or a compound may not have completed a full evaluation, if such data are available they should be MIABE-compliant

How does MIABE differ from the 'guidelines to authors' provided by journals?

- Compared with standard author guidance, the MIABE guidelines are much more specific. The guidelines explicitly list every piece of information that should be provided, leaving nothing open to interpretation
- The MIABE checklist does not address data quality in any form; such judgments are the province of reviewers (who will be better equipped to form such judgments if they are provided with a MIABE-compliant data set)
- The MIABE checklist does not recommend the use of any particular protocol and only requests that each protocol is described in full detail

What tools are available to enable me to use MIABE?

Normally, the MIABE guidelines would be consulted as a manuscript is being prepared, and compliance requires no dedicated tool. However, if you wish to prepare a database submission or use existing data in a database, several tools are available, including:

- Dedicated controlled vocabularies such as the Human Proteomics Organization Proteomics Standards Initiative-Molecular Interactions (HUPO PSI-MI) controlled vocabulary to enable data annotation
- Both XML and MITAB (tab-delineated) schemas in which data can be downloaded from one resource and moved into a second. The use of such formats enables the merging of separate data sets, including with protein interaction networks or pathway data, and enables access to additional visualization and analysis tools such as Cytoscape
- A common web service (the Proteomics Standard Initiative Common Query Interface; PSICQUIC), which enables data search over multiple resources

How can I have input into future versions of MIABE?

Although it is our intention to keep all modules stable for as long as possible (at least one year per version), updates and extensions will be required as new techniques are developed and existing ones evolve. The MIABE guidelines can be accessed through the Minimum Information for Biological and Biomedical Investigations (MIBBI) website and contact e-mails will be maintained. The module has been developed, in part, by the Molecular Interactions workgroup of the HUPO-PSI-MI worktrack, which has an annual Spring workshop that is open to anyone to attend, plenary meetings at the HUPO congress, a mailing list and regular phone conferences on data formats and controlled vocabulary updates.

We also intend to regularly review the MIABE standard through the European Bioinformatics Institute (EBI) Industry Programme — which is a forum for interaction between the European Molecular Biology Laboratory (EMBL)-EBI and the industrial life science research sector — and to further disseminate its message through the small and medium-sized enterprise (SME) Support Forum meetings (see the EBI Industry Programme website). Discussions are also welcome on the ChEMBL-og website.

Mechanism-of-action studies. Other studies that contribute to the understanding of the mechanism of action of a particular molecule should be reported using the appropriate part (or parts) of the MIABE guidelines. Such data could include the stoichiometry of binding or the mechanism by which the compound binds to its target.

Pharmacokinetic studies. In cases in which pharmacokinetic studies have been carried out, they should be reported in full detail. Calculated or predicted data should be clearly differentiated from experimentally verified data, and any algorithms that were used to determine the calculated values should be described.

Biomarkers. During the past decade, various biomarkers — for example, levels of specific plasma proteins or microarray expression profiles — of disease processes and drug responses have been increasingly investigated in preclinical and clinical studies of potential drugs. For example, in preclinical studies such biomarkers may be used to aid the selection of the best candidate compounds to advance to clinical trials, and in the clinic they can be used to stratify patients with subtypes of a disease based on their likelihood of responding to a particular drug. If biomarker studies are included, they should conform to the same standards as any other bioactive entity, and be fully and unambiguously described. As the amount of data in this area increases and compounds move into preclinical and clinical development, it will be necessary to revisit and update existing standards to ensure that these data can be fully described.

Conclusion

The main aim of the MIABE document is to support the core informatics-related aspects of the long-term bioactive molecule discovery process, ensuring that the data generated from previous expensive and time-consuming projects are not lost but can be harvested and built on in the future. By following such guidelines, authors can produce a useful, richly annotated data set about the molecule or series of molecules that they are describing, which will allow effective data retrieval by anyone wishing to reproduce or analyse the results, or undertake comparative and/or predictive studies.

The publication of the MIABE guidelines is recognized by the authors of this article as being only a first step in the long process of making all data on bioactive molecules readily available in a single common format, with full and consistent data description. Key challenges in achieving this goal will initially include the spreading of this message to the widely disparate group of workers in this field, some of whom have commercial interests that constrain their data exchange policies and format usage. Finding a single professional body to represent this group has also proved to be an issue, and as a result existing standards bodies have been utilized to supply an independent platform from which to launch this standard. These bodies will also be important in maintaining the MIABE checklist, to incorporate new technologies and advances into the procedure (see BOX 1). The endorsement and adoption of this standard by public database resources will also be crucial, and many such resources have had a key role in developing the MIABE checklist so far.

It is hoped that more groups will recognize the value of putting such data into the public domain to become part of an ever-increasing knowledge bank on bioactive molecules. This would aid the development of improved

molecules with a well-understood mode of action, high safety and environmental compatibility to target an expanded druggable genome in an increasing number of species.

Sandra Orchard, Dominic Clark, Anna Gaulton, Christoph Steinbeck, Henning Hermjakob, John Overington and Janet Thornton are at the European Molecular Biology Laboratory (EMBL)-European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK.

Bissan Al-Lazikani is at the Cancer Research UK Cancer Therapeutics Unit, The Institute of Cancer Research, 15 Cotswold Road, Belmont, Surrey SM2 5NG, UK.

Steve Bryant is at the National Center for Biotechnology Information, National Library of Medicine, US National Institutes of Health, Bldg. 38A, 8600 Rockville Pike, Bethesda, Maryland 20894, USA.

Elizabeth Calder, Ian Dix and Nick Lynch are at AstraZeneca R&D, R&D Information, Alderley Park, Macclesfield SK10 4TG, UK.

Ola Engkvist is at DECS Computational Sciences, AstraZeneca R&D Mölndal, SE-431 83 Mölndal, Sweden.

Mark Forster is at Syngenta R&D Information Systems, International Research Centre, Jealott's Hill, Bracknell, Berkshire RG42 6EY, UK.

Michael Gilson is at the University of California San Diego, Skaggs School of Pharmacy and Pharmaceutical Sciences, La Jolla, California 92093, USA.

Robert Glen and Peter Murray-Rust are at the Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK.

Martin Grigorov and Elena Lo Piparo are at the Nestlé Research Center, PO BOX 44, CH-1000 Lausanne 26, Switzerland.

Kim Hammond-Kosack is at Rothamsted Research, Harpenden, Hertfordshire, AL5 2JQ, UK.

Lee Harland is at Pfizer, Ramsgate Road, Sandwich, Kent CT13 9NJ, UK.

Lee Harland is also at Connected Discovery Ltd, 27 Old Gloucester Street, London WC1N 3AX, UK.

Andrew Hopkins is at Biological Chemistry and Drug Discovery, College of Life Sciences, University of Dundee, Dundee DD1 5EH, UK.

Christopher Larminie and Romeena K. Mann are at GlaxoSmithKline, Stevenage, Hertfordshire, SG1 2NY, UK.

Christopher Southan is at ChrisDS Consulting, Göteborg, 42166, Sweden.

David Wishart is at the Departments of Computing Science & Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2E9,Canada.

Correspondence to S.O.
e-mail: orchard@ebi.ac.uk

doi:10.1038/nrd3503

1. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nature Rev. Drug Discov.* **1**, 727–730 (2002).
2. Hajduk, P. J., Huth, J. R., & Tse, C. Predicting protein druggability. *Drug Discov. Today* **10**, 1675–1682 (2005).
3. Schmidtke, P. & Barril, X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.* **53**, 5858–5867 (2010).
4. Halgren, T. A. Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.* **49**, 377–389 (2009).
5. Zhang, Z. & Grigorov, M. G. Similarity networks of protein binding sites. *Proteins* **62**, 470–478 (2006).
6. Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? *Nature Rev. Drug Discov.* **5**, 993–996 (2006).
7. The UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.* **39**, D214–D219 (2010).
8. Berriman, M. *et al.* The genome of the blood fluke *Schistosoma mansoni*. *Nature* **460**, 352–358 (2009).
9. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME) — toward standards for microarray data. *Nature Genet.* **29**, 365–371 (2001).
10. Degtyarenko, K. *et al.* ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* **36**, D344–D350 (2008).
11. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **39**, D38–D51(2011).
12. Seiler, K. P. *et al.* ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.* **36**, D351–D359 (2008).
13. Pederson, F., de Brujin, J., Munn, S. & van Leeuwen, K. Assessment of additional testing needs under REACH. Effects of (Q)SAR, risk based testing and voluntary industry activity. European Commission report EUR 20863 (2003).
14. Vizcaino, J. A., Martens, L., Hermjakob, H., Julian, R. K. & Paton, N. W. The PSI formal document process and its implementation on the PSI website. *Proteomics* **7**, 2355–2357 (2007).
15. Yeung, B. K. *et al.* Spirotetrahydro β-carbolines (spiroindolones): a new class of potent and orally efficacious compounds for the treatment of malaria. *J. Med. Chem.* **53**, 5155–5164 (2010).
16. Kerrien, S. *et al.* Broadening the horizon — level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.* **5**, 44 (2007).
17. Aranda, B. *et al.* PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nature Methods* **8**, 528–529 (2011).
18. Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39**, D691–D697 (2011).
19. Taylor, C. F. *et al.* Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature Biotech.* **26**, 889–896 (2008).
20. Orchard, S. *et al.* The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nature Biotech.* **25**, 894–898 (2007).
21. Bourbeillon, J. *et al.* Minimum Information about a protein affinity reagent (MIAPAR). *Nature Biotech.* **28**, 650–653 (2010).
22. Schuffenhauer, A. *et al.* An ontology for pharmaceutical ligands and its application for *in silico* screening and library design. *J. Chem. Inf. Comput. Sci.* **42**, 947–955 (2002).
23. Qi, D. *et al.* An ontology for description of drug discovery investigations. *J. Integr. Bioinform.* 25 Mar 2010 (doi:10.2390/biecoll-jib-2010-126).
24. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
25. Stein, S. E., Heller, S. R. & Tchekhovskoi, D. An open standard for chemical structure representation: the IUPAC chemical identifier. *Proceedings of the 2003 International Chemical Information Conference (Nîmes), Infonortics,* 131–143 (2003).
26. Inglese, J., Shamu, C. E. & Guy, R. K. Reporting data from high-throughput screening of small-molecule libraries. *Nature Chem. Biol.* **3**, 438–441 (2007).
27. Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.* **38**, D331–D335 (2010).

Acknowledgements
This work has, in part, been funded by the European Commission under the Proteomics Standards Initiative and International Molecular Exchange (PSIMEx), contract number FP7-HEALTH-2007-223411.

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION

ChEMBL database: https://www.ebi.ac.uk/chembl
ChEMBL-og website: http://chembl.blogspot.com
Cytoscape: http://www.cytoscape.org
Disease Ontology Community Wiki: http://diseaseontology.sourceforge.net
EBI Industry Programme: http://www.ebi.ac.uk/industry
EU-Openscreen project: https://www.ebi.ac.uk/chembl/eu-openscreen/project
HUPO Proteomics Standards Initiative website: http://www.psidev.info
HUPO PSI-MI XML 2.5 documentation: http://www.psidev.info/index.php?q=node/60
InchI Trust: http://www.inchi-trust.org
Innovative Medicines Initiative: http://www.imi.europa.eu
IUPAC website: http://www.iupac.org/web/ins/2000-025-1-800
MIACA: www.mibbi.org/index.php/Projects/MIACA
MIACA homepage: http://miaca.sourceforge.net
MIBBI Foundry: http://mibbi.org/index.php/MIBBI_foundry
MIBBI portal: http://www.mibbi.org/index.php/MIBBI_portal
MIBBI website: http://www.mibbi.org
MIMIx guidelines: http://www.mibbi.org/index.php/Projects/MIMIx
Molecular Interactions workgroup: http://www.psidev.info/index.php?q=node/31
Molecular Libraries Program: http://mli.nih.gov/mli/mlpcn
Ontology Lookup Service: http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=MI
Open Biological and Biomedical Ontologies: http://www.obofoundry.org
PSICQUIC web service: http://code.google.com/p/psicquic
STRENDA guidelines: http://www.mibbi.org/index.php/Projects/STRENDA

SUPPLEMENTARY INFORMATION

See online article: S1 (table) | S2 (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

ONLINE CORRESPONDENCE

Nature Reviews Drug Discovery publishes items of correspondence online. Such contributions are published at the discretion of the Editors and may be subject to peer review. Correspondence should be a scholarly comment on a specific article that has been published in the journal. To view correspondence, please go to our home page at http://www.nature.com/reviews/nrd and select the link from the current table of contents.

The following correspondence has recently been published:

Believe it or not: how much can we rely on published data on potential drug targets?

Florian Prinz, Thomas Schlange and Khusru Asadullah

doi:10.1038/nrd3532-c1

This correspondence relates to the article:

Phase II failures: 2008–2010

John Arrowsmith

Nature Rev. Drug Discov. **10**, 328–329 (2011)