



Review

The BRENDa enzyme information system—From a database to an expert system



I. Schomburg, L. Jeske, M. Ulbrich, S. Placzek, A. Chang, D. Schomburg*

Technische Universität Braunschweig, Braunschweig Integrated Centre of Systems Biology (BRICS), Rebenring 56, 38106 Braunschweig, Germany

ARTICLE INFO

Keywords:

Enzyme database
Metabolic pathways
Enzyme-catalysed reactions
Enzyme kinetics
Enzymes and diseases
Enzyme-ligand interaction

ABSTRACT

Enzymes, representing the largest and by far most complex group of proteins, play an essential role in all processes of life, including metabolism, gene expression, cell division, the immune system, and others. Their function, also connected to most diseases or stress control makes them interesting targets for research and applications in biotechnology, medical treatments, or diagnosis. Their functional parameters and other properties are collected, integrated, and made available to the scientific community in the BRAunschweig ENzyme Database (BRENDA). In the last 30 years BRENDA has developed into one of the most highly used biological databases worldwide. The data contents, the process of data acquisition, data integration and control, the ways to access the data, and visualizations provided by the website are described and discussed.

1. Introduction

1.1. Relevance of enzymes

Enzymes are essential to almost all processes of life and vital to industrial biotechnology or medical diagnostics. 30–40% of all genes encode enzymes. They accelerate chemical reactions by up to 16 orders of magnitude, allow for precisely coordinated metabolic pathways within cells, and are indispensable when it comes to defence against pathogens and other processes. Many of them are highly specific, while others are less so. The functions of enzymes are dependent on many characteristics, such as their sequence, three-dimensional structure, stability, and their interactions with other molecules (Schomburg and Schomburg, 2016).

In the late 80s it became obvious that the molecular or cellular function of enzymes or their role in disease processes depends on a large number of different properties including e.g. their substrate specificity, their stability, their location, and many others and that these properties vary between different organisms although the general catalysed biochemical reaction is the same on first view.

These data – different for example from sequence and 3D structure information – is not directly available but hidden in millions of publications and has to be extracted and converted to a structural form. This only can be handled in a flexible database system and made accessible via a website.

1.2. History

The collection of enzyme-related data from the scientific literature started in 1987 at the German National Research Centre for Biotechnology in Braunschweig (GBF). From 1996–2007 it was continued at the University of Cologne and returned to Braunschweig to the department of Bioinformatics and Systems Biology at the Technische Universität Braunschweig. It is now located at the Systems Biology Centre Braunschweig (BRICS) at <http://www.brenda-enzymes.org>.

The initial data consisted of a structured data compilation in textual and numeric form, with one dataset for each enzyme class. These were published from 1990 on in a series of 19 books covering ~3,000 enzyme classes (Schomburg et al., 1990–1998). A second edition consisting of 49 vols covering ~4,900 enzyme classes was published from 2001 to 2009 (Schomburg and Schomburg, 2001–2006).

Concurrently with the books an internet presentation of the database system was created starting from a first text-based query system in 1999 (Schomburg et al., 2000; Schomburg et al., 2002). Since then it was developed into an elaborate comprehensive enzyme information system covering more than 370,000 enzymes from 6,300 enzyme classes. It includes a wealth of enzyme-related information by combining the manually annotated functional data with genomic sequences, enzyme structures, and computed data. The additional data retrieved by text mining algorithms are aiming at providing a complete set of literature for classified enzymes. Calculated values and information on the catalysed reactions and compounds modulating enzyme activities including their molecular structures add to the comprehensive

* Corresponding author.

E-mail address: [D.Schomburg](mailto:D.Schomburg@tu-braunschweig.de).

BRENDA platform of these days (Chang et al., 2009; Chang et al., 2015; Placzek et al., 2017). Since 2015 BRENDA is a member of the German Network for Bioinformatics Infrastructure (<https://www.denbi.de/>) (see Section 5.3.).

1.3. Data import & integration

Enzymes are classified by the enzyme task force of the IUBMB (International Union of Biochemistry and Molecular Biology), according to the catalysed reactions (McDonald et al., 2009; McDonald and Tipton, 2014). A consequence of this concept is that a certain name designates not a single enzyme protein but a group of proteins with the same catalytic property. Each single enzyme class, denominated by a four-digit identifier, the EC number, may be populated with only one or thousands of different enzymes from a large number of different organisms, their common characteristics being that all are able to catalyse the specified reaction.

BRENDA data are either extracted manually from the primary literature or obtained from integration of data from other sources or acquired by text mining (and disclosed as such). Each single entry is connected to a reference, covering the literature between 1939 and 2016. It is also connected to an organism, and where available to a strain and a sequence identifier for the enzyme-protein. The manually annotated data cover ~60 categories grouped into ‘Nomenclature’, ‘Enzyme-Ligand Interactions’, ‘Functional Parameters’, ‘Organism-related Information’, ‘Enzyme Structure’, ‘Molecular Properties’, ‘Bibliography/Links/Disease’. Each EC class is updated periodically, data from selected new references are annotated and included (see Table 1).

Table 1
Number of entries in selected BRENDA information fields. The numbers refer to basic information on the given values for a specific enzyme in a given organism.

Overview	
EC Classes	7,100
Proteins	84,000
Enzyme Names	104,000
Ligands	211,000
References	133,000
Application	13,000
Engineering	80,000
Organism-related Information	
Organism	11,000
Source Tissue	97,000
Localization	33,000
Reaction & Specificity	
Substrates/Products	930,000
Cofactors	26,000
Inhibitors	220,000
Activating Compounds	28,000
Metals & Ions	38,000
Functional Parameters	
K _M Values	137,000
Turnover Numbers	64,000
K _{cat} /K _M Values	21,000
Specific Activity	48,000
IC50 Values	51,000
K _i Values	38,000
P _i Values	5,000
pH Optima & Ranges	53,000
Temperature Optima & Ranges	32,000
Enzyme Structure	
Posttranslational Modifications	6,000
Quaternary Structure	35,000
Isolation & Preparation	
Stability Values (pH, Temperature, Oxidation, Storage, Organic Solvent)	49,000
Purification	33,000
Cloned	33,000
Renatured	1,000

1.3.1. Converting hidden unstructured data to structured data: manual data import and extension

The creation of the BRENDA information core consists of the following steps:

- 1 Selection of relevant papers from PubMed (NCBI Resource Coordinators, 2015) and Scopus (Burnham, 2006)
- 2 Extraction of information items from the selected papers
- 3 Quality control of extracted information items
- 4 Integration into the internal BRENDA database
- 5 Addition of structural information for new enzyme ligands
- 6 Data integration of external data from websites
- 7 Precalculation of e.g. genome annotations, location predictions, statistics etc.
- 8 Integration into one main database
- 9 Biannual release to users

Independently from the database update, the interface, query engine, and other software programs are continuously optimised.

1.3.1.1. Enzyme information. The selection of relevant papers from PubMed and Scopus is a very important step. 2.6 Million papers are listed in PubMed under the MeSH (Medical Subject Headings, Rogers, 1963) term “enzymes”, more than 100,000 of which appeared in 2015. This means that only a tiny fraction of all papers can be exploited. The goal of the selection process is to identify those papers that give a wide overview on the enzyme in question. Depending on the state of knowledge for a certain enzyme between 1 and 700 papers are included in the knowledge base. The literature search is based on a combined search in PubMed and Scopus using terms related to the data categories in BRENDA, i.e. enzyme names, enzyme function, reactions (substrates/products), and occurrence (organism, tissues etc.).

The first step in this process is the identification of the different names used for each enzyme. Although the IUBMB publishes an “accepted name” for each enzyme class currently ~79,000 synonyms are listed for the 6,300 EC classes. These manually annotated synonyms are essential for the identification of the relevant literature and also for the text mining process (vide infra). Some enzymes are more often cited with a synonym than with the “accepted name”. A prominent example is ribulose bisphosphate carboxylase (EC 4.1.1.39) which is commonly known as RuBisCo.

After the manual selection and analysis steps the relevant literature is annotated by highly qualified scientists.

The current BRENDA release (2017.1) covers 6,300 active enzyme classes (EC numbers). BRENDA also contains “preliminary BRENDA-supplied EC numbers”. These numbers are designated with a “B” in the fourth position of the EC number. These are enzymes found by manual literature annotation and differ substantially in substrate specificity and reaction from all currently classified enzymes. Sometimes they are enzymes closing a gap in a metabolic pathway or enzymes of a newly detected pathway in a specific organism. These enzymes have to be classified within the hierarchical EC system and are waiting for the final decision of the IUBMB Enzyme Commission.

1.3.1.2. Enzyme ligand data. Information on compounds interacting with enzymes such as substrates and products, cofactors, inhibiting and activating substances constitute a major part of BRENDA (Scheer et al., 2011). These can be “small molecules”, or macromolecules such as proteins, polynucleotides, and polysaccharides. In the biochemical literature, however, the use of compound names is highly inconsistent. The systematic IUPAC nomenclature is rarely used, instead trivial names, abbreviations, and acronyms are most often found (IUPAC, 2017). Currently BRENDA stores 211,000 ligand names. The 124,000 different small molecule ligands are found under 170,000 different

names in the literature. Among them are compounds which are referred to in the literature under up to 70 different names. 20,500 compounds possess at least two names.

Very often trivial names and acronyms are not unique and are even used for different compounds. An exact mapping of synonymous ligands is only possible via a comparison of their chemical structures. This means that for all ligands the chemical structure has to be included, in most cases hand drawn and stored as “Molfiles”, InChI codes (Heller et al., 2015), and figures.

1.3.1.3. Quality control. The data manually extracted from each paper are channelled through an elaborated control workflow starting with hundreds of different computer-based checks controlling the formal accuracy and the internal consistency of the data. This is followed by manual controls by two internal scientists. This process guarantees a high quality of the data before they are further processed for the final BRENDa database release.

2. Import into and compilation of the BRENDa database

The databases for the BRENDa website consist of numerous tables representing the manually annotated data as well as external or automatically generated data. The full SQL main database originating from the manual core, data integration and data mining processes, and calculated data is optimised for fast query performance and presently consists of 550 tables.

2.1. Data integration and data mining

2.1.1. Sequence information

Information on sequences is retrieved from UniProt (The UniProt consortium, 2017): The data are used to supplement the BRENDa information, to calculate additional information (e.g. transmembrane helices), and to calculate the BrEPS enzyme sequence patterns (Bannert et al., 2010).

2.1.2. Enzyme 3D structure information

The coordinates of all known enzyme 3D-structures are imported from the Protein Data Bank (Berman et al., 2000) and analysed. Information of active or binding sites, glycosylation sites, disulfide bridges etc. are extracted and presented to the users in an interactive 3D molecular graphics visualization.

2.1.3. Pathway information

BRENDa metabolic pathways are extended by pathways from KEGG (Kanehisa et al., 2016) and MetaCyc (Caspi et al., 2014) in order to provide the possibility to compare an enzyme's metabolic context in all three databases on the enzyme summary pages (see chapter 3.2.1).

2.1.4. Genomes

The genomes from the EMBL-EBI (Li et al., 2015) build the base for the Genome Explorer. The extracted genomes, UniProt accessions, and positions are supplemented by protein names and EC numbers derived from UniProt, from predicted EC numbers doing a BLAST search (Altschul et al., 1990) in UniRef (Suzek et al., 2015), and KEGG metabolic pathways. For every gene the neighbourhood consisting of three genes on each side is calculated.

2.1.5. Data converted for hierarchical browsing

Some of the imported data are converted from a tabular form to a tree-like representation for a hierarchical analysis and accession of the data. This includes:

The EC Explorer (see chapter 3.1.1.3), a tree-like view of all currently approved enzyme classes, is created based on data provided by the IUBMB.

The Taxonomy Tree allows the user to search and browse for

organisms or enzymes along the taxonomic hierarchical tree. The results lead to the specific enzyme data and the enzyme summary pages. The tree is based on the NCBI Taxonomy database (Federhen, 2012) which is the main resource for organisms in BRENDa and supplemented with microbial strain denominations. Presently BRENDa contains 11,420 different organisms. The organisms and where available strains are linked to the respective NCBI pages and to the Bacterial Diversity Metadatabase BacDive (Söhngen et al., 2016). A small number of organisms, which are not stored at the NCBI, are verified by other databases or by the original literature.

SCOPe (Fox et al., 2014) and CATH (Sillitoe et al., 2015) are two hierarchical classification schemes for protein 3D structures. They are implemented in BRENDa to afford the identification of the folding class of an enzyme or the identification of enzymes with a similar folding in different levels. For instance, the NAD(P)-binding Rossmann-like domain is associated with 186 different EC numbers.

The disease branch of the MeSH ontology is also part of the ontology section. It forms the basis for the hierarchical access to the part of BRENDa which covers enzyme/disease relationships in DRENDA, the Disease Related ENzyme DAtabase (Söhngen et al., 2011; Schomburg et al., 2013).

In Table 2 the amount of data of all external sources is summarized.

2.1.6. The BRENDa tissue ontology (BTO)

With the increasing knowledge about the occurrence of enzymes in tissues and organs it was recognised that many isoenzymes show tissue-specific properties. To access e.g. all enzymes occurring in liver or in plant roots the development of an organism-independent ontology for tissues, organs, anatomical structures, plant parts, cell types, cell lines, and cell cultures proved to be essential. Based on the rules and formats of the Gene Ontology Consortium (The Gene Ontology Consortium, 2015) ~6,000 terms (January 2017) are organized as a directed acyclic graph. Most of the terms are supplemented with information on synonyms, on their origin, and definitions. The BTO allows the user to distinguish and allocate enzymes due to their occurrence and provides an easy-to-use platform to perform simple or advanced searches (see Fig. 1). All entries within the ontology are directly connected to the comprehensive enzyme data in BRENDa (Gremse et al., 2011).

Not only for BRENDa but also in the field of proteomics the BTO is one of the important and recommended ontologies. It is meanwhile widely used in the life science community to evaluate tissue samples (Harhay et al., 2010) or to map tissue-specific gene annotations to the BTO (Greene et al., 2015; Santos et al., 2015). The controlled vocabulary of the BTO is often used as a repository/source to annotate, to interpret, or to compare their corresponding data (Wittig et al., 2014). The URSA platform (Unveiling RNA Sample Annotation) incorporates the BTO to predict tissue/cell-type signals in a gene expression sample (Lee et al., 2013). The ProteomeXchange Consortium recommends to use the BTO as a reference during the submission procedure (ProteomeXchange Submission Tutorial).

2.1.7. Other ontologies

In 2015 the human anatomy atlas CAVEman was linked to the BTO

Table 2
Overview of external data sources integrated into BRENDa.

Data Source	Data Type	Amount
IUBMB	Enzymes	6,896
NCBI	Organisms	1,537,142
UniProt	Sequences	65,930,418
EMBL EBI	Genomes	24,638
UniRef	Accessions	54,163,854
PDB	3D structures	63,464
KEGG/MetaCyc	Pathways	7,600/6,812

Details for blood vessel

BRENDA blood vessel

BTO (BRENDA Tissue Ontology) ID
BTO:0001102

BTO:0001102 is linked to 344 enzymes:
1.1.1.1 Show enzyme

Definition
Any of the vessels through which blood circulates in the body

References
1. From Merriam-Webster's Online Dictionary at www.Merriam-Webster.com:
<http://www.m-w.com/cgi-bin/dictionary?book=Dictionary&va=blood+vessel>

Condensed Tree View

- └ tissues, cell types and enzyme sources
 - └ animal
 - └ whole body
 - └ cardiovascular system
 - └ vascular system
 - └ blood vessel

Tree view

tissues, cell types and enzyme sources

- └ tissues, cell types and enzyme sources
- └ animal
 - └ whole body
 - └ adult stem cell
 - └ body wall
 - └ cardiovascular system
 - └ diaphragm
 - └ heart
 - └ pericardial mesothelium
 - └ pericardium
 - └ vascular system
 - └ arterial system
 - └ blood vessel
 - └ cerebrovascular system

Search term: blood vessel

EC Number ▼▲	Recommended Name ▼▲	Source Tissue ▼▲
1.1.1.21	aldehyde reductase	blood vessel
1.1.1.49	glucose-6-phosphate dehydrogenase (NADP+)	blood vessel
1.1.1.146	11beta-hydroxysteroid dehydrogenase	blood vessel
1.1.1.188	prostaglandin-F synthase	blood vessel
1.2.1.3	aldehyde dehydrogenase (NAD+)	blood vessel
1.4.3.4	monoamine oxidase	blood vessel
1.6.2.4	NADPH-hemoprotein reductase	blood vessel
1.6.3.1	NAD(P)H oxidase (H2O2-forming)	blood vessel
1.8.5.1	glutathione dehydrogenase (ascorbate)	blood vessel
1.14.13.8	flavin-containing monooxygenase	blood vessel

◀◀ Results 1 - 10 of 43 ▶▶

Fig. 1. View on the term *blood vessel* in the BRENDA Tissue Ontology and the connected EC classes.

to connect human enzymes of BRENDA to the detailed ontology for body structures, organs, and tissues based on the nomenclature of 'Terminologia Anatomica' (FCAT, 1998). The terms provide direct links to the functional enzyme data, their properties, specificity, and medical relevance to give insights to biomedical issues and disease-related information in the human body.

In addition a number of other, more specialised ontologies are integrated and connected to the appropriate BRENDA enzymes. This includes ontologies like human, mouse, *Drosophila*, fungal anatomies, plant structures, and the Gene ontology.

2.2. Text mining

The manual data annotation from the primary literature is limited to about 10,000 articles per year and (due to budget limitations) cannot cover the complete literature. The BRENDA database, however, tries to give a complete literature overview for each enzyme class providing basic information such as the occurrence in different organisms. This information is retrieved by text mining of literature abstracts from PubMed. Due to the license conditions of the publishers this process is presently limited to an analysis of PubMed titles and abstracts and therefore restricted to information expected to be mentioned therein. It is based on extended dictionaries and complex algorithms. The fragmentation, analysis, and quality control requires weeks of computer time on a high performance cluster.

The – presently – 26 million titles and 16 million abstracts from PubMed are split into words and sentences. The dictionaries for the text mining comprising around 9,000 source tissues, 80,000 enzyme names, 2.2 million organism synonyms, 170,000 disease terms, 130 million substance names and 4,000 kinetic expressions are automatically compiled. In a manual evaluation ambiguous or false terms are eliminated.

The first part of the subsequent text mining (Table 3) identifies co-

Table 3
Data retrieved by text mining.

	BRENDA manual	AMENDA	FRENDA	UniProt
References	137,000	1,700,000	2,500,000	
Organism Information	83,000	400,000	700,000	Swiss-Prot 185,000 TrEMBL 4.3 million
Tissue Information	91,000	320,000		
Subcellular Localization	33,000	80,000		

occurrences of these terms in the processed titles and abstracts. The following steps differ depending on the specific purpose.

2.2.1. Occurrence of enzymes in organisms/tissues/subcellular localizations

The FRENDA database (Full Reference of ENzyme DAta) is based on all co-occurrences of enzyme names and organisms in the titles and abstracts and includes all combinations (Barthelmes et al., 2007).

In a subsequent step this is further processed to predict the important enzyme/organism combination for each paper and to extend this by analysing organ/tissue and subcellular localization information (AMENDA data). The post-processing includes automatic validations based on the BRENDA tissue ontology and manually created exclusion lists for AMENDA.

This analysis shows that within the manual BRENDA core about 6% of the papers mentioning enzymes are processed, and hereby between 25% and 40% of the relevant information is included.

2.2.2. Kinetic data

In a highly complex process kinetic data are extracted from titles and abstracts. This includes a position-specific analysis of numeric

values connected to terms used for enzyme kinetics, enzyme information, organism, and names for chemical compounds. Unfortunately kinetic information is rarely given in abstracts, so the number of values obtained in this way is limited. 11,000 papers with kinetic data were identified and 13,000 kinetic values were extracted. This adds to the 410,000 kinetic data obtained by the manual processing from full papers.

2.2.3. Disease data

As enzyme function or dysfunction/deficiency is often connected to diseases or used for diagnosis, the inclusion of disease information is essential. Papers in this area are not suitable for the extraction of structured information as described above. For this purpose a text mining process was developed that identifies papers covering the role of enzymes with respect to diseases from different points of view. A first co-occurrence-based step for the identification of relevant papers and enzyme-related diseases is followed by a classification step based on a technique used in supervised learning and data mining as well as classification, the so-called “support vector machine”.

In this way 770,000 papers are identified to cover enzyme/disease relations. They are classified to give either information on causal interaction between enzyme and disease (450,000, 350,000 with the highest confidence score), papers describing enzyme-based treatment (330,000) or diagnosis (380,000), and other subjects. Several scores based on precision or recall-preference can be chosen by the users.

2,000 enzyme classes were found in papers describing causal interactions, with protein kinases, monooxygenases, and ubiquitin transferases mentioned most often. 3,600 disease terms are mentioned in connection to enzymes, with neoplasms/carcinoma and infections mentioned most often.

2.3. Prediction of data based on calculation

2.3.1. Trans-membrane or membrane association of enzymes

For the prediction of trans-membrane or membrane association of enzymes TMHMM is used (Krogh et al., 2001). The user does not only get predictions on the membrane-associated sequences of individual enzymes but can ask for all enzyme classes characterised by a certain number of trans-membrane helices. For the 20,526,372 predicted transmembrane proteins the location of the N-terminus and the start and end points are determined.

2.3.2. Genome annotation, function prediction of gene products – EnzymeDetector

The rapidly rising number of sequenced genomes provides an invaluable source of information on the occurrence of enzymes in organisms, provided a reliable sequence-based prediction of enzyme function is available. Since an analysis (Quester and Schomburg, 2011) showed that the standard genome annotation hosts give different function predictions for more than two thirds of all enzymes, an aggregated view on the most popular prediction hosts (NCBI, KEGG, UniProt, PATRIC (Gillespie et al., 2011)), Pfam-HMMs (Finn et al., 2016; Finn et al., 2015) was integrated into the BRENDA service and combined with the manually annotated data in BRENDA, a recent BLAST search, the text mining data from AMENDA, and an own, orthogonal pattern-based enzyme function prediction. The EnzymeDetector presently provides a comprehensive collection of genome-wide enzyme function predictions for more than 5,000 bacterial and 200 archaeal genomes including their plasmids (at <http://edbs.tu-bs.de> and linked at BRENDA).

The combination and aggregation of gene function prediction from different sources and obtained by different methods allows the assignment of a confidence score to each prediction, depending on the agreement of the different sources. The user can then modify the default scores and decide on the cutoff score, which could be the assumption that 30 percent of the proteins are enzymes, a number proven for

Escherichia coli (UniProt) and generally assumed for prokaryotic organisms.

2.4. The final database

In the last step, information from the described sources are combined. Frequently requested specific combinations of parameters are pre-compiled. The ligand molfiles are converted into figures for the visualization of ligands and reactions (see chapter 3.2.6) and fingerprints are calculated for the substructure search (see chapter 3.1.1.2).

3. The BRENDA host

Owing to a complex system of several servers, load balancers, and a version control system the BRENDA host www.brenda-enzymes.org experienced 0 min downtime in 2015 and 2016 and is even accessible during updates. The query engines and web pages are updated constantly in order to optimise user access which is often based on user feedback. The complete database is converted from the internal normal form to a form optimised for short reaction times.

Recently, there have been major revisions of the BRENDA web pages: A new and modern interface has been implemented for the entry page. The enzyme and ligand summary pages have been completely revised including new print functions and options to filter the content, to sort all tables, or to hide data fields. Furthermore, the client-side Java applications for the substructure search (chapter 3.1.1.2) and the visualization of 3D structures (chapter 3.2.8) have been substituted by JavaScript applications.

3.1. Data access – query engine

The data in BRENDA are – different from for example the data contained in PubMed or UniProt – stored in different formats. They are of alphabetical, numerical, structural types like chemical structures, reactions, pathways, ontologies, and the mentioned tree structures. Often combinations of information are requested by the users like “enzymes that occur in a certain organism and display a certain substrate specificity”. In fact most of the users just want to have a quick overview of the full information on a certain enzyme class, others, however, are trying to identify an enzyme for a certain application that requires a specific combination of properties like a pH or temperature specificity and may have a broad substrate specificity.

The BRENDA query engine and webpages are designed to cover the needs of many different kinds of users with as few “clicks” as possible.

3.1.1. Entry page

The entry page is divided into three parts. The header presents options to quickly navigate to different search tools (“go to”), to return to the entry page (“HOME”), or to switch to a page giving access to all different search fields (“Classic view”). Moreover, the header gives access to a login page, a form to open a ticket, a search history, or an overview of all enzymes.

In the main part of the page the user can simply enter a search term in a Google search manner and choose the desired search field (see Fig. 2). For the default search field “Enzyme or Ligand” a life search offering instant results during the input of the search term is implemented.

Additionally, six tiles represent the six different more elaborated search strategies:

3.1.1.1. Text-based queries. Text-based queries provide access to alphanumerical fields. They cover the full-text search, the advanced search, and the possibility to explore enzymes/disease relations. The advanced search allows the user to combine different search criteria. The full-text search includes the commentaries and identifies the desired search term in almost all parts of the database. As an example

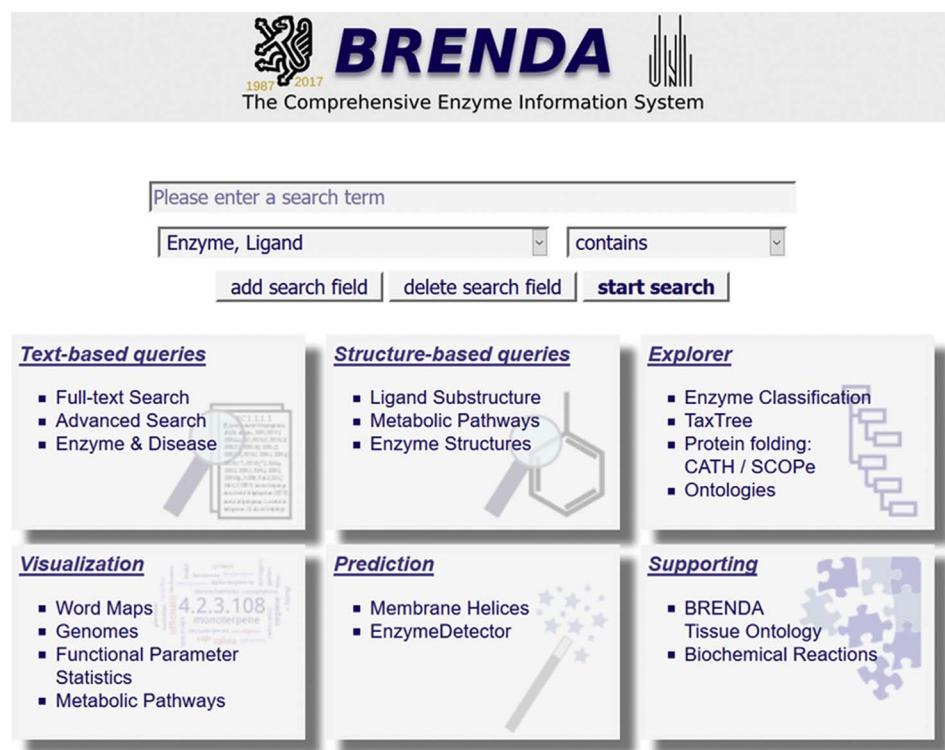


Fig. 2. The BRENDA entry page giving access to various search options.

the term “wastewater” is found in the fields application, cloned, inhibitors, reference titles, source tissues, specific activity, and substrates (in most cases in the commentaries).

3.1.1.2. Structure-based queries. As mentioned, biochemical compounds are referred to in the literature often by a large number of highly different terms which renders an alphanumerical search for compound names very inefficient. Often information is wanted for a class of compounds that share a certain substructure. The highly efficient query strategy is based on the use of a molecular editor (JavaScript based JSME Editor (Bienfait and Ertl, 2013)) that allows the user to enter a chemical structure. The BRENDA engine identifies this structural element within the 155,000 ligand structures in BRENDA by a highly optimised substructure search algorithm. The substructure search is done in two steps: A pre-selection is carried out via fingerprint scans, the remaining structures are then checked for subgraph matchings. The result is a list of ligands with all synonyms, their roles, and structure diagrams. The search can be limited to a certain role of the ligand interaction with the enzyme.

3.1.1.3. Tree-browsing and queries. The enzyme classification, the taxonomy tree, the protein folding, and several ontologies can be explored by browsing a tree-like representation of the data or searching for a specific term or synonym and give information how many enzymes are found in the particular branch.

The EC Explorer gives an overview on the EC classes. A short summary and a link to the detailed information for each EC class is displayed. Additionally, the reaction diagram can be displayed and the known sequences and PDB identifiers can be downloaded.

The TaxTree Explorer allows the user to search and browse for organisms or enzymes along the taxonomic hierarchical tree. The results lead to the enzyme summary pages, the sequences, and the metabolic pathways known for the selected organism.

The Ontology Explorer gives access to 35 different ontologies including the BRENDA Tissue Ontology (BTO), CAVEman, Structural Classification of Proteins (SCOPe), the Protein Structure Classification

Database (CATH), and Medical Subject Headings (MeSH). All ontologies are connected to BRENDA enzymes and localizations where possible.

3.1.1.4. Visualization. A number of visualizations like word maps, pathway maps, genome explorer, catalysed reactions, and protein 3D structures have been compiled that permit a quick access to data (compare 3.2).

3.1.1.5. Calculated parameters. In the fifth category predicted parameters like transmembrane helices or the EnzymeDetector can be accessed.

3.1.1.6. Supporting information. Finally, a sixth category extends BRENDA's query methods by supporting tools like the BTO and integrated BRENDA, KEGG, MetaCyc, and SABIO-RK (Wittig et al., 2012) biochemical reactions (BKMS-react, (Lang et al., 2011)).

A footer focuses on the essential supporting information and completes the better user experience.

3.2. Data representation, visualization, and data export

3.2.1. Enzyme summary page

The enzyme summary page is the most frequently called page in BRENDA. It provides the full and often huge amount of various data of an enzyme. The information about an enzyme is divided into the main categories enzyme nomenclature, enzyme-ligand interactions, functional parameters, organism related information, enzyme structure data, molecular properties, diseases, references, and links to other databases. With the search form at the top of the page, the user can restrict the displayed data by choosing a certain reference, a UniProt accession, an organism, or a list of organisms. Additionally, the results can be extended by text mining hits. For many enzymes with more than 5 references a word map is displayed with related enzyme-specific terms from PubMed titles and abstracts (see chapter 3.2.4). The different entries are complemented by links to the diverse range of

BRENDA tools. Organisms are connected to the TaxTree Explorer, reactions with reaction diagrams, UniProt IDs to a sequence view, and the references to the reference summary page.

3.2.2. Ligand summary page

The ligand summary page pools all available data of a ligand in BRENDA. This view is divided into four sections. The basic ligand information includes a graphical representation of the molecule, the chemical sum formula, the InChI key, a list of synonyms, and a link to the respective metabolic pathways. The second part describes the roles as enzyme ligand in enzyme-catalysed reactions and the function as activator or inhibitor. All entries are connected to an EC number and the corresponding BRENDA literature reference. The third section comprises enzyme kinetic parameters including inhibition constants. All literature references of the molecule and hyperlinks to ChEBI (Hastings et al., 2013) and PubChem (Kim et al., 2016) appear at the end of the ligand summary page.

3.2.3. Reference summary page

Each information unit in BRENDA is connected to a literature reference, displayed directly next to the parameter. A click on the reference ID opens a page with the PubMed title and entry of this reference (if available in PubMed) and a compilation of all values extracted from this paper.

3.2.4. Word maps

Many users want to have a quick overview on the special role, application, or properties of a certain enzyme. Due to the high amount of information about an enzyme stored in BRENDA this is not always easily possible by scanning the enzyme summary page. To provide the user with an initial overview of scientific topics and facts published in context with this enzyme, word maps are developed for 3,584 EC classes where more than 5 references are cited in PubMed. They connect the enzyme with a collection of specific terms extracted from paper abstracts and titles.

Based on the reference lists for each EC number obtained by the text mining process a frequency distribution of terms is generated for words with the same stem. After execution of diverse filtering methods, for instance, the removal of common words occurring in many different scientific texts, the terms are ranked according to their enzyme specificity. Maximally 50 of them are then arranged in a word map displayed on the enzyme summary page and search result pages. The font size of a term relates to its specificity. Additionally, terms are colour-coded into the categories *enzyme*, *source tissue*, *localization*, *disease*, *organism*, *application*, *ligand*. The different categories are linked to the respective BRENDA query page to get more information. For example, in the word map of the enzyme acetylcholinesterase the ligand donepezil (in light blue), used as medication to treat Alzheimer's disease, is connected with the ligand summary page (see Fig. 3).

3.2.5. Pathway maps

Since 2016 BRENDA has integrated maps showing enzymes in their metabolic context providing an alternate intuitive access to enzymes and metabolites. The current full map contains 9,774 nodes for metabolites and enzymes and 10,198 edges which combine enzymes and reactants to reactions. The reactions are organized in 154 pathways of different sizes. Almost all reactions are connected to EC classes. The current version represents 1,674 EC classes of which 165 are incompletely classified, either because the enzyme for the respective metabolic step has not yet been described in the literature or the reaction is spontaneous and does not need a catalyst.

The BRENDA pathway entry page shows a multicoloured overview map. The pathways are coloured according to their metabolic function as *central and energy metabolism*, *lipid metabolism*, *amino acid metabolism*, *nucleotide and cofactor metabolism*, *carbohydrate metabolism*, *fermentation and other catabolism*, and *xenobiotics and secondary metabolism*. On

mouseover the multicoloured areas show the name of the pathways and a click directly leads to the corresponding maps. On the left side the user finds a menu with an alphabetical list of pathways and various search options.

Search for specific enzymes or metabolites can be performed via the name or the EC number or parts thereof. Entering a search term will result in a partly coloured overview map which highlights the pathways where the enzyme or metabolite of interest is involved. For the convenience of the users the colours can be changed individually. A combined search for a ligand, an enzyme, and an organism genus is also possible. The organism search can be performed on two sets of database information. The default mode reverts to the manually verified BRENDA data. It is possible to expand the displayed data by including the BRENDA text mining data from the AMENDA and FRENDNA data subsets.

Frequently, the user looks for an overview on the metabolic capacity of an organism or of a taxonomic range of organisms. This can be obtained by a combination of an organism search and the visualization of taxonomic information. The taxonomic range is indicated by a colouring scheme. The further up in the taxonomic range the lighter the colour becomes. Within each map the nodes are linked to the BRENDA data. Clicking on an EC number leads to the enzyme summary page, while clicking on a metabolite leads to the ligand summary page respectively.

Figs. 4 and 5 show an example how taxonomic information combined with pathway occurrence can be retrieved for the bacterium *Staphylococcus aureus*. Thereby, all taxonomic information about enzymes starting from the organism of interest and ending with the bacterial level is highlighted with decreasing colour intensity. The user can select a taxonomic level by switching intermediate levels on or off. Finally, this result map also contains the coverage of highlighted nodes related to the number of all enzyme nodes in a pathway via mouseover. This feature distinguishes the BRENDA pathways from the other two main pathway databases KEGG and MetaCyc. These platforms offer either an overview map for all organisms or species-specific maps. There are also pathway databases which focus on a single organism. Examples are SMPDB (The Small Molecule Pathway Database) for human enzymes (Jewison et al., 2014). The Yeast Metabolome Database (YMDB) is a manually curated database for small molecule metabolites found in or produced by *Saccharomyces cerevisiae* (Ramirez-Gaona et al., 2017).

3.2.6. Biochemical reactions

Enzyme-catalysed reactions are represented in structure diagrams consisting of the pictures of the involved ligands created during the BRENDA update (see chapter 1.3). The reaction diagrams are available on every result page where a reaction is mentioned.

3.2.7. Genome explorer

The Genome Explorer visualizes enzymes in their genomic context (i.e. three genes before or after its position). The user can either choose an available genome from a drop-down list or perform a search with an organism, within a taxonomic range, for an EC number, or for a specific protein. The query results in a list of genomes. After one or more genomes are chosen the genomes are visualized. The Genome Explorer offers the possibility to move along the genome, to zoom in or out, or to click on an EC number to open detailed enzyme information.

3.2.8. Protein 3D structures

For the visualization of 3D structures the Jsmol viewer is used (<http://www.rcsb.org/pdb/home/home.do#Category-visualize>). The viewer illustrates the protein folding as well as disulfide bonds, active sites, binding sites or regions, and glycosylation sites. Additionally, links to the visualization of the protonated enzyme (Bietz et al., 2014) or enzyme pockets Volkamer et al., 2012) are displayed.



Fig. 3. Word map showing enzyme-specific terms from PubMed titles and abstracts for EC 3.1.1.7 – acetylcholinesterase catalysing the initial step in the degradation of the neurotransmitter acetylcholine.

3.2.9. Statistics

The Functional Parameter Statistics allows the user to visualize and compare the full value distribution of numerical parameters stored in BRENDA including all kinetic parameters as well as pH profile or temperature optima of organism classes.

3.2.10. SBML, SOAP output

There are three ways to download data from the BRENDA web portal, either interactively e.g. as CSV files or program-initiated as SBML or via SOAP. For the SBML download an organism has to be specified. The search can be refined by using further search criteria like the EC number. Additionally, a SOAP web service is offered for academic users. After a successful registration the BRENDA data can be downloaded using several methods, e.g. "getKmValue" for extracting all K_m -value entities in BRENDA.

3.2.11. BKMS-react

For the development of metabolic models or for an assessment of the metabolic capacities of a certain organism, complete information on biochemical reactions is required. However, the reactions stored in the BRENDA enzyme database, and in the KEGG and MetaCyc databases, or the kinetic values in SABIO-RK are different due to different priorities (compare chapter 5). In order to provide the BRENDA user with a full overview on all reactions stored in the databases they are combined into the integrated reaction database BKMS-react containing currently 65,367 unique reactions and 6,644 EC numbers originating from BRENDA, KEGG, MetaCyc, and SABIO-RK. Identical and redundant reactions are identified by aligning metabolites on the basis of molecule name and structure comparisons.

The information about reactions is complemented by EC numbers, pathway information, hyperlinks to the various databases, commentaries as well as a stoichiometry check. Stoichiometric unbalanced reactions are marked and missing atoms at the reactant and product

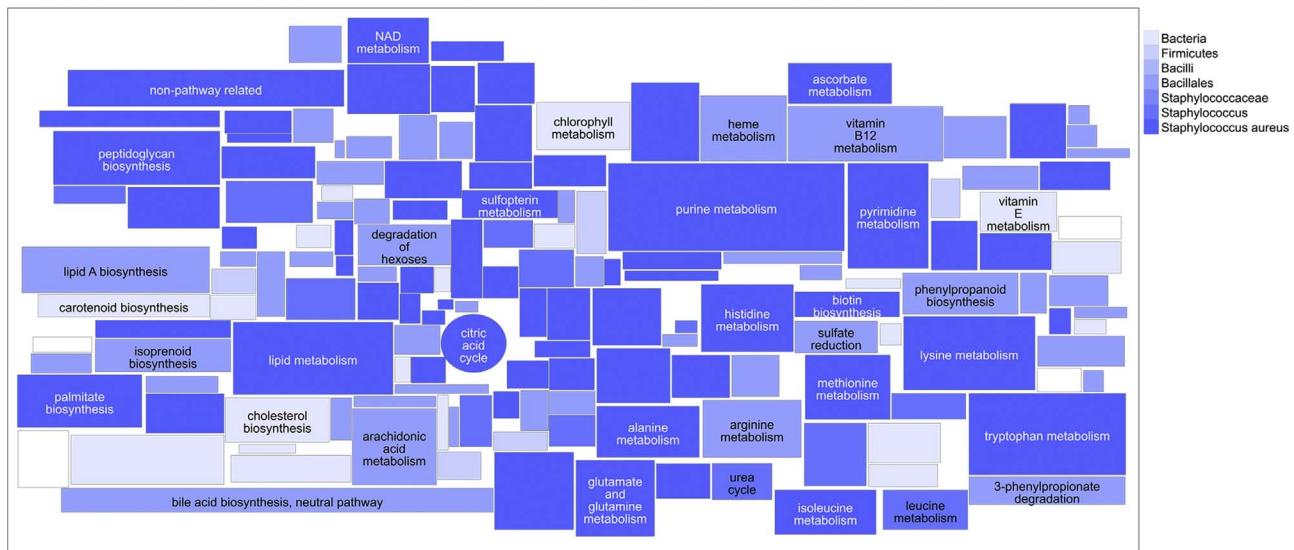


Fig. 4. Metabolic pathways with manually curated enzyme data for the bacterium *Staphylococcus aureus* shown in dark blue. Pathways with enzymes from taxonomically related organisms up to the bacterial level are indicated by lighter colours.

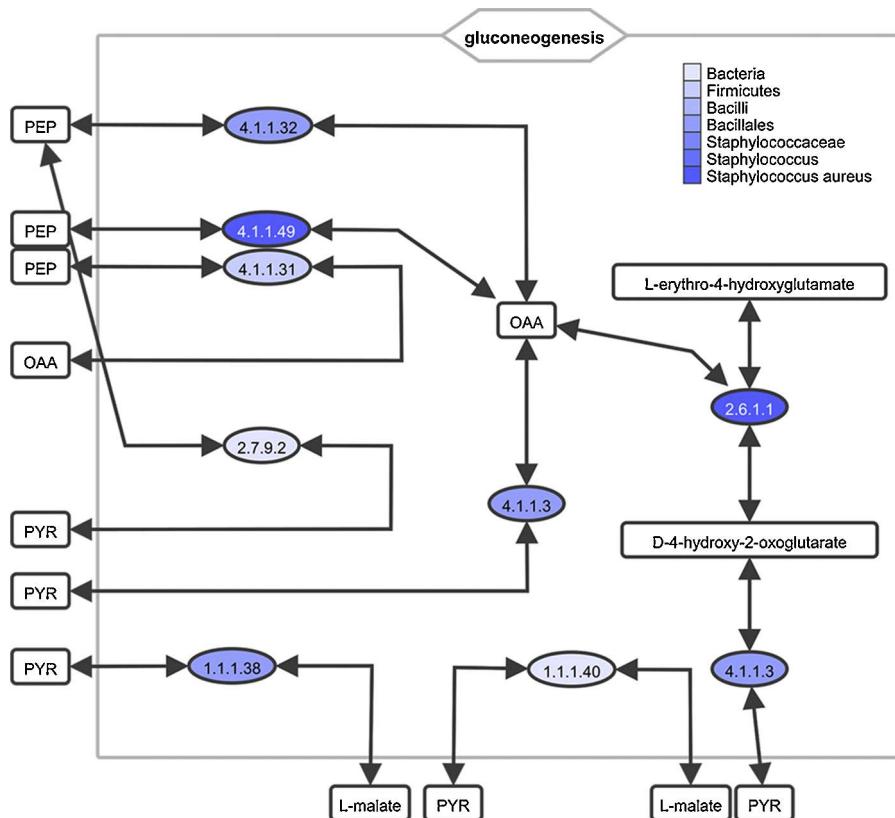


Fig. 5. Enzymes from *Staphylococcus aureus* manually curated for gluconeogenesis. Enzymes from taxonomically related organisms up to the bacterial level are indicated by lighter colours. The nodes for enzymes and metabolites are linked to the corresponding BRENDA information. Further information on co-metabolites and cofactors can be displayed on demand.

side are indicated excluding entries where only H₂O or protons are absent.

44,681 reactions from BRENDA (68% of all unique reactions), 3,191 from KEGG, 7,596 from MetaCyc and 444 from SABIO-RK occur only in the specified databases. 1,348 reactions can be found in all four databases and 3,542 in BRENDA, KEGG, and MetaCyc (see Fig. 6).

4. User support

4.1. User feedback

The rapid advance of science in the biosciences has a strong influence on the needs and expectations of scientists accessing biological databases. A continuous contact to users via the website, on

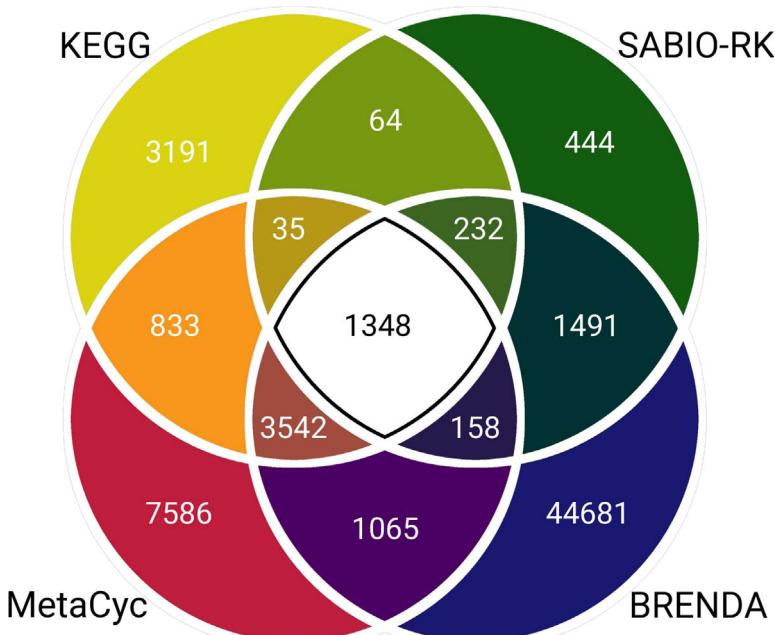


Fig. 6. Distribution of unique reactions between BRENDA, KEGG, MetaCyc, and SABIO-RK in BKMS-react.

conferences, teaching courses, and cooperation are the essential requirements for the development of service-oriented bioinformation infrastructures. Based on these contacts and surveys, the data contents in BRENDA have been continuously enlarged and the website has been refurbished a number of times. User contacts will continue to be an important factor also in the future development of BRENDA. A ticket system was established to provide a web interface for the users to send their questions, to suggest new enzymes, to point to inconsistencies in BRENDA, or to ask for an integration of specific data into the database.

4.2. Tutorials – trainings

BRENDA offers diverse options for the users to learn more about the content, the scope and the different tools, and how to use the enzyme information system. The user can study the online training materials, which contain an introduction, details on search functions (quick, advanced or full-text, ligand search) or the ontology search and browse functions (EC Explorer, BTO, TaxTree), sequence and genomes searches, and the visualization tools (word maps, pathway maps). Since 2015 these topics are also shown in video tutorials which are particularly helpful for beginners. In the past training sessions were held in several European universities. Currently in the context of the de.NBI project 2-day training courses are offered. These are organized to present the database, to initiate discussions, and to demonstrate practical applications.

5. Discussion

BRENDA was started in the second half of the 1980s at the German national lab for biotechnology when enzymes started to be exploited in larger scale as biocatalysts for the production of complex compounds and it became necessary to identify suitable enzymes with the correct reaction, substrate specificity, and stability. Starting with a rather moderate aim to produce a kind of one-page overview for the then known EC classes which was intended to be published in form of a book, the initial setup proved to be highly future-proof as the form of a highly structured and controlled data representation was chosen which could – much later – easily be converted into a relational database. The development of the internet which was able to transport larger amounts of data only since the mid 90s offered the possibility to present the data via the network, first at the EBI, since 1999 directly at the University of Cologne.

Meanwhile BRENDA has developed into one of the most highly used bioinformation infrastructures worldwide, despite the fact that the number of scientists within the BRENDA team has been rather low (1–5 over the years, now 4). This was only possible because the BRENDA team always had a close interaction with the different user communities working with enzymes and quickly responded to the scientific progress in the biosciences, especially gene, protein, and genome sequencing, the progress in computing power, allowing elaborated data integration and text-mining procedures and a continuous improvement of the query engine and the computing setup.

5.1. BRENDA as a unique resource for the biosciences

Given the wide range of enzyme-specific parameters and the coverage of all kinds of enzymes from all organisms and the size of the information it offers, BRENDA is unique. This is reflected by the large number of users accessing the database (up to 100,000 users per month).

Of course there are other databases providing enzyme and metabolic information, data on specific aspects of enzyme, covering a specific class of enzymes or giving pathway information. The most important ones will be discussed briefly.

The annotated sequence database UniProt contains basic functional information on sequenced genes and proteins, including enzymes. KEGG (Kyoto Encyclopedia of Genes and Genomes)

comprises genomic data, enzyme nomenclature information, and organism-specific metabolic maps. Literature references are restricted to those obtained from the IUBMB enzyme classification database. MetaCyc's main feature are metabolic pathways and the involved enzymes, genes, and physiological substrates and cofactors. Inhibitors are not included, nor kinetic or other enzyme data. Enzyme protein 3D structures are deposited in and provided by the Protein Data Bank PDB. Other databases such as FunTree (Furnham et al., 2012) which uses the CATH classification to annotate and analyse structure-based families of proteins are based on the crystallographic data deposited therein.

Reactome (Fabregat et al., 2016) is a database focused on humans. Among other pathways (signalling, transport, gene expression, etc.) it describes metabolic processes in the cell. Its hierarchical organization largely follows that of the Gene Ontology. The MEROPS database (Rawlings et al., 2016) is a special compendium of peptidases. The classification of families and subfamilies is based on the similarity of protein sequences. Carbohydrate-modifying enzymes are characterised in the CAZY database (Lombard et al., 2014). The classification into families and subfamilies is mainly based on protein sequence similarity and phylogenetic analysis.

In addition, there are numerous other mostly smaller databases highlighting specific aspects of enzymes. Enzyme catalytic mechanisms for 321 EC classes are covered in the MACIE (Holliday et al., 2012) and EzCatDB (Nagano et al., 2015) which covers 871 different enzymes based on the type of reaction. Detailed kinetic data for a selection of ~17,300 enzyme-catalysed reactions can be found in SABIO-RK. Enzyme-catalysed reactions are collected in RHEA (Morgat et al., 2017), a resource of manually curated biochemical reactions. It includes the reactions from the IUBMB enzyme nomenclature websites (4,794) plus 4,479 additional reactions including reactions from the Swiss Lipids Knowledgebase and spontaneous reactions occurring in biological systems. (Aimo et al., 2015). The EAWAG-BBD (Ellis and Wackett, 2012) database, now located in Switzerland, focuses on the degradation of xenobiotic compounds by microorganisms. It covers ~1,500 reactions and 543 microorganism strains.

5.2. Usage of BRENDA – application cases

The BRENDA website is used by 80,000–100,000 users per month. The large majority of users is mainly interested in one or a small number of enzymes. They usually do not cite BRENDA in their papers or mention the URL, nor the paper. This is different when scientists use a larger part of the information stored in BRENDA. Unfortunately, the name BRENDA was chosen before the era of the internet, so, because it is identical to a personal first name it is impossible to identify the number of webpages linking to BRENDA. Presently 1500 papers cite BRENDA publications.

Analysis of the titles of these papers gives an impression on the range of applications. The short selection shown in Fig. 7 is not intended to rate the applications but is selected from those papers that are themselves cited far beyond 100 times and should show highly diverse applications. In addition to enzymology and biochemistry the major fields of BRENDA usage include biotechnology, medical and pharmaceutical research, systems biology and modelling, evolution, and many others. Not mentioned here are papers describing other databases that depend on data from BRENDA, like Transfac (Kaplin et al., 2016), ChEMBL (Bento et al., 2014), the Golm Metabolome Database (Hummel et al., 2013), the Catalytic Site Atlas (Furnham et al., 2014), and many others.

5.3. BRENDA in de.NBI bioinformatics infrastructure

BRENDA covers the whole complex field of enzyme-related information in de.NBI (German Network for Bioinformatics Infrastructure) and is part of the service center for *Biological Data*. Besides the

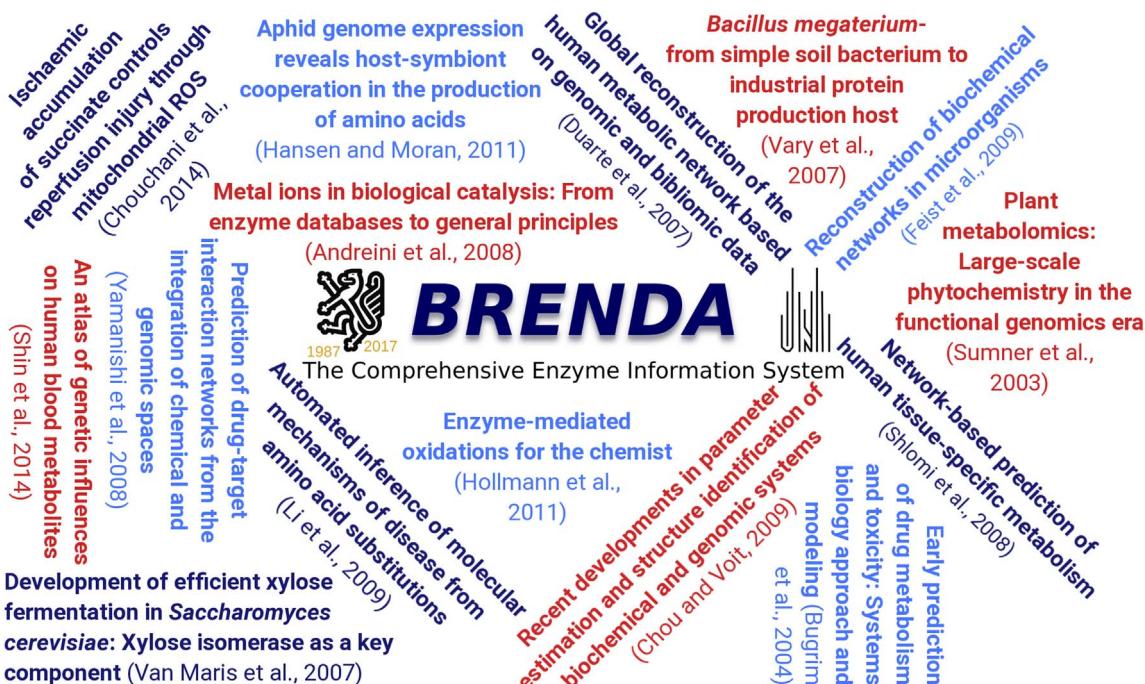


Fig. 7. A selection of publication titles representing the various application cases of BRENDa (Andreini et al., 2008; Bugrim et al., 2004; Chou and Voit, 2009; Chouchani et al., 2014; Duarte et al., 2007; Feist et al., 2009; Hansen and Moran, 2011; Hollmann et al., 2011; Li et al., 2009; Shin et al., 2014; Shlomi et al., 2008; Sumner et al., 2003; Van Maris et al., 2007; Vary et al., 2007; Yamanishi et al., 2008).

contribution to the Special Interest Groups *training & education* and *service & service monitoring*, a collaboration with several project partners from de.NBI exists to exchange and integrate data.

BKMS-react, a non-redundant database for biochemical reactions in BRENDA (see chapter 3.2.1.1), has recently been extended by 3,785 reactions from SABIO-RK. This database originally consisted of reactions from BRENDA, KEGG, and MetaCyc. SBML files from SABIO-RK were parsed and the extracted reactions were integrated by aligning substrates and products on the basis of their chemical structure.

Moreover, the BRENDA websites include various links to de.NBI project partners. The EC classes in BRENDA are linked to reaction kinetic data of SABIO-RK. Furthermore, there are links to BacDive via strain information of bacteria and archaea in the TaxTree Explorer (see chapter 3.1.1.3). Links to DogSiteScorer (Volkamer et al., 2012), a software to detect potential binding pockets, and ProToss (Bietz et al., 2014), a fully automated hydrogen prediction tool for protein-ligand complexes, are inserted into the 3D structure search result pages in BRENDA. In a further cooperation with the Rarey group from the Universität Hamburg, it could be possible to create a new web view in BRENDA addressing enzyme ligand interactions.

BRENDA gives high priority to the user education within the framework of the de.NBI training. A training workshop for users is organized every year to give an overview about the variety of tools and all possibilities to run quick and advanced enzyme searches. Additionally, constantly updated handouts, hands-on exercises, and training videos are provided (see chapter 4.2).

BRENDA is freely available for academic users and educational purposes. Commercial users need to obtain a license. The license fees are essential for the maintenance and development of BRENDA.

Funding

This work was funded by the German Federal Ministry of Education and Research (BMBF) [grant numbers 031A539D, 01KX1235, 0316188F, and 031L0078G] and by the Niedersächsisches Ministerium für Wissenschaft und Kultur [74ZN1122].

References

- Aimo, L., Liechti, R., Hyka-Nouspikel, N., Niknejad, A., Gleizes, A., Gotz, L., Kuznetsov, D., David, F.P., van der Goot, F.G., Riezman, H., Bougueret, L., Xenarios, I., Bridge, A., 2015. The SwissLipids knowledgebase for lipid biology. *Bioinformatics* 31, 2860–2866.

Altshul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Andreini, C., Bertini, I., Cavallaro, G., Holliday, G.L., Thornton, J.M., 2008. Metal ions in biological catalysis: from enzyme databases to general principles. *J. Biol. Inorg. Chem.* 13, 1205–1218.

Bannert, C., Welfle, A., aus dem Spring, C., Schomburg, D., 2010. BrEPS: a flexible and automatic protocol to compute enzyme-specific sequence profiles for functional annotation. *BMC Bioinf.* 11, 589.

Barthelmes, J., Ebeling, C., Chang, A., Schomburg, I., Schomburg, D., 2007. BRENDA: AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Res.* 35, D511–D514.

Bento, A.P., Gaulton, A., Hersey, A., Bellis, L.J., Chambers, J., Davies, M., Krüger, F.A., Light, Y., Mak, L., McGinchey, S., Nowotka, M., Papadatos, G., Santos, R., Overington, J.P., 2014. The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* 42, D1083–D1090.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. *Nucleic Acids Res.* 28, 235–242.

Bienfait, B., Ertl, P., 2013. JSME: a free molecule editor in JavaScript. *J. Cheminform.* 5, 24.

Bietz, S., Urbaczek, S., Schulz, B., Rarey, M., 2014. Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. *J. Cheminform.* 6, 12.

Bugrim, A., Nikolskaya, T., Nikolsky, Y., 2004. Early prediction of drug metabolism and toxicity: systems biology approach and modeling. *Drug Discov. Today* 9, 127–135.

Burnham, J.F., 2006. Scopus database: a review. *Biomed. Digit. Libr.* 3, 1.

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C.A., Holland, T.A., Keseler, I.M., Kothari, A., Kubo, A., Krumannacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D.S., Weerasinghe, D., Zhang, P., Karp, P.D., 2014. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* 42, D459–D471.

Chang, A., Scheer, M., Grote, A., Schomburg, I., Schomburg, D., 2009. BRENDA: AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *Nucleic Acids Res.* 37, D588–D592.

Chang, A., Schomburg, I., Placzek, S., Jeske, L., Ulbrich, M., Xiao, M., Sensen, C.W., Schomburg, D., 2015. BRENDA in 2015: exciting developments in its 25th year of existence. *Nucleic Acids Res.* 43, D439–D446.

Chou, I.-C., Voit, E.O., 2009. Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Math. Biosci.* 219, 57–83.

Chouchani, E.T., Pell, V.R., Gaude, E., Aksentijevic, D., Sundier, S.Y., Robb, E.L., Logan, A., Nadtochiy, S.M., Ord, E.N.J., Smith, A.C., Eyyasu, F., Shirley, R., Hu, C.-H., Dare, A.J., James, A.M., Rogatti, S., Hartley, R.C., Eaton, S., Costa, A.S.H., Brookes, P.S., Davidson, S.M., Duchen, M.R., Saeb-Parsy, K., Shattock, M.J., Robinson, A.J., Work, L.M., Frezza, C., Krieg, T., Murphy, M.P., 2014. Ischaemic accumulation of succinate controls reperfusion injury through mitochondrial ROS. *Nature* 515, 431–435.

Duarte, N.C., Becker, S.A., Jamshidi, N., Thiele, I., Mo, M.L., Vo, T.D., Srivas, R., Palsson, B.O., 2007. Metabolic modeling of Escherichia coli leads to novel insights of the central metabolism. *Nature* 450, 328–332.

- B.O., 2007. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. U. S. A.* 104, 1777–1782.
- Ellis, L.B., Wackett, L.P., 2012. Use of the University of Minnesota Biocatalysis/Biodegradation Database for study of microbial degradation. *Microb. Inf. Exp.* 2, 1.
- FCAT, 1998. Terminologia Anatomica International Anatomical Terminology. Thieme, New York.
- Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., Matthews, L., May, B., Milacic, M., Rothfels, K., Shamovsky, V., Webber, M., Weiser, J., Williams, M., Wu, G., Stein, L., Hermjakob, H., DEustachio, P., 2016. The reactome pathway knowledgebase. *Nucleic Acids Res.* 44, D481–D487.
- Federhen, S., 2012. The NCBI taxonomy database. *Nucleic Acids Res.* 40, D136–D143.
- Feist, A.M., Herrgard, M.J., Thiele, I., Reed, J.L., Palsson, B.O., 2009. Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.* 7, 129–143.
- Finn, R.D., Clements, J., Arndt, W., Miller, B.L., Wheeler, T.J., Schreiber, F., Bateman, A., Eddy, S.R., 2015. HMMER web server: 2015 update. *Nucleic Acids Res.* 43, W30–W38.
- Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., Bateman, A., 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, D279–D285.
- Fox, N.K., Brenner, S.E., Chandonia, J.M., 2014. SCOPe: structural classification of proteins-extended: integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* 42, D304–309.
- Furnham, N., Sillitoe, I., Holliday, G.L., Cuff, A.L., Rahman, S.A., Laskowski, R.A., Orengo, C.A., Thornton, J.M., 2012. FunTree: a resource for exploring the functional evolution of structurally defined enzyme superfamilies. *Nucleic Acids Res.* 40, D776–D782.
- Furnham, N., Holliday, G.L., de Beer, T.A., Jacobson, J.O., Pearson, W.R., Thornton, J.M., 2014. The catalytic site atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.* 42, D485–D489.
- Gillespie, J.J., Wattam, A.R., Cammer, S.A., Gabbard, J.L., Shukla, M.P., Dalay, O., Driscoll, T., Hix, D., Mane, S.P., Mao, C., Nordberg, E.K., Scott, M., Schulman, J.R., Snyder, E.E., Sullivan, D.E., Wang, C., Warren, A., Williams, K.P., Xue, T., Yoo, H.S., Zhang, C., Zhang, Y., Will, R., Kenyon, R.W., Sobral, B.W., 2011. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect. Immun.* 79, 4286–4298.
- Greene, C.S., Krishnan, A., Wong, A.K., Ricciotti, E., Zelaya, R.A., Himmelstein, D.S., Zhang, R., Hartmann, B.M., Zaslavsky, E., Sealfon, S.C., Chasman, D.I., FitzGerald, G.A., Dolinski, K., Grosser, T., Troyanskaya, O.G., 2015. Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* 47, 569–576.
- Gremse, M., Chang, A., Schomburg, I., Grote, A., Scheer, M., Ebeling, C., Schomburg, D., 2011. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res.* 39, D507–D513.
- Hansen, A.K., Moran, N.A., 2011. Aphid genome expression reveals host-symbiont co-operation in the production of amino acids. *Proc. Natl. Acad. Sci. U. S. A.* 108, 2849–2854.
- Harhay, G.P., Smith, T.P., Alexander, L.J., Haudenschild, C.D., Keele, J.W., Matukumalli, L.K., Schroeder, S.G., Van Tassel, C.P., Gresham, C.R., Bridges, S.M., Burgess, S.C., Sonstegard, T.S., 2010. An atlas of bovine gene expression reveals novel distinctive tissue characteristics and evidence for improving genome annotation. *Genome Biol.* 11, R102.
- Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., Steinbeck, C., 2013. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.* 41, D456–D463.
- Heller, S.R., McNaught, A., Pletnev, I., Stein, S., Tchekhovskoi, D., 2015. InChI, the IUPAC international chemical identifier. *J. Cheminform.* 7, 23.
- Holliday, G.L., Andreini, C., Fischer, J.D., Rahman, S.A., Almonacid, D.E., Williams, S.T., Pearson, W.R., 2012. MACiE: exploring the diversity of biochemical reactions. *Nucleic Acids Res.* 40, D783–D789.
- Hollmann, F., Arends, I.W.C.E., Buehler, K., Schallmey, A., Bühler, B., 2011. Enzyme-mediated oxidations for the chemist. *Green Chem.* 13, 226–265.
- Hummel, J., Strehmel, N., Bölling, C., Schmidt, S., Walther, D., Kopka, J., 2013. Mass spectral search and analysis using the golem metabolome database. *The Handbook of Plant Metabolomics*. Wiley-VCH Verlag GmbH & Co. KGaA, pp. 321–343.
- IUPAC website: <https://iupac.org/>.
- Jewison, T., Su, Y., Disfany, F.M., Liang, Y., Knox, C., Maciejewski, A., Poelzer, J., Huynh, J., Zhou, Y., Arndt, D., Djoumbou, Y., Liu, Y., Deng, L., Guo, A.C., Han, B., Pon, A., Wilson, M., Rafatnia, S., Liu, P., Wishart, D.S., 2014. SMPDB 2.0: big improvements to the small molecule pathway database. *Nucleic Acids Res.* 42, D478–D484.
- Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M., 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44, D457–D462.
- Kaplin, A., Krull, M., Lakshman, K., Matys, V., Lewicki, B., Hogan, J.D., 2016. Establishing and validating regulatory regions for variant annotation and expression analysis. *BMC Genomics* 17 (2), 393.
- Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B.A., Wang, J., Yu, B., Zhang, J., Bryant, S.H., 2016. PubChem substance and compound databases. *Nucleic Acids Res.* 44, D1202–D1213.
- Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580.
- Lang, M., Stelzer, M., Schomburg, D., 2011. BKM-react, an integrated biochemical reaction database. *BMC Biochem.* 12, 42.
- Lee, Y.S., Krishnan, A., Zhu, Q., Troyanskaya, O.G., 2013. Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies. *Bioinformatics* 29, 3036–3044.
- Li, B., Krishnan, V.G., Mort, M.E., Xin, F., Kamati, K.K., Cooper, D.N., Mooney, S.D., Radivojac, P., 2009. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* 25, 2744–2750.
- Li, W., Cowley, A., Uludag, M., Gur, T., McWilliam, H., Squizzato, S., Park, Y.M., Buso, N., Lopez, R., 2015. The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res.* 43, W580–W584.
- Lombard, V., Golacanda Ramulu, H., Drula, E., Coutombard, V., Golacanda Ramulu, H., Drula, E., Coutinho, P.M., Henrissat, B., 2014. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42, D490–D495.
- McDonald, A.G., Tipton, K.F., 2014. Fifty-five years of Enzyme classification: advances and difficulties. *FEBS J.* 281, 583–592.
- McDonald, A.G., Boyce, S., Tipton, K.F., 2009. ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Res.* 37, D593–D597.
- Morgat, A., Lombardot, T., Axelsen, K.B., Aimo, L., Niknejad, A., Hyka-Nouspikel, N., Couder, E., Pozzato, M., Pagni, M., Moretti, S., Rosanoff, S., Omwubiko, J., Bougueret, L., Xenarios, I., Redaschi, N., Bridge, A., 2017. Updates in Rhea – an expert curated resource of biochemical reactions. *Nucleic Acids Res.* 45, D415–D418.
- Resource Coordinators, N.C.B.I., 2015. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 44, D7–D19.
- Nagano, N., Nakayama, N., Ikeda, K., Fukui, M., Yokota, K., Doi, T., Kato, T., Tomii, K., 2015. EzCatDB: the enzyme reaction database, 2015 update. *Nucleic Acids Res.* 43, D453–D458.
- Placzek, S., Schomburg, I., Chang, A., Jeske, L., Ulrich, M., Tillack, J., Schomburg, D., 2017. BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res.* 45, D380–D388.
- ProteomeXchange Submission Tutorial: http://genesis.ugent.be/files/costore/practicals_peptide_and_protein_identification_tutorial/source/software/PX_Submission_ProteomeXchange_Submission_Tutorial.pdf.
- Quester, S., Schomburg, I., 2011. EnzymeDetector: an integrated enzyme function prediction tool and database. *BMC Bioinf.* 12, 376.
- Ramirez-Gaona, M., Marcu, A., Pon, A., Guo, A.C., Sajed, T., Wishart, N.A., Karu, N., Djoumbou Feunang, Y., Arndt, D., Wishart, D.S., 2017. YMDB 2.0: a significantly expanded version of the yeast metabolome database. *Nucleic Acids Res.* 45, D440–D445.
- Rawlings, N.D., Barrett, A.J., Finn, R., 2016. Twenty years of the MEROPS database of proteolytic enzymes: their substrates and inhibitors. *Nucleic Acids Res.* 44, D343–D350.
- Rogers, F.B., 1963. Medical subject headings. *Bull. Med. Libr. Assoc.* 51, 114–116.
- Söhngen, C., Chang, A., Schomburg, D., 2011. Development of a classification scheme for disease-related enzyme information. *BMC Bioinf.* 12, 329.
- Söhngen, C., Podstawka, A., Bunk, B., Gleim, D., Vetcinina, A., Reimer, L.C., Ebeling, C., Pendarowski, C., Overmann, J., 2016. BacDive-the bacterial diversity metadata-base in 2016. *Nucleic Acids Res.* 44, D581–D585.
- Santos, A., Tsafou, K., Stoltz, C., Pletscher-Frankild, S., O'Donoghue, S.I., Jensen, L.J., 2015. Comprehensive comparison of large-scale tissue expression datasets. *Peer. J.* 3, e1054.
- Scheer, M., Grote, A., Chang, A., Schomburg, I., Munaretto, C., Rother, M., Söhngen, C., Stelzer, M., Thiele, J., Schomburg, D., 2011. BRENDA: the enzyme information system in 2011. *Nucleic Acids Res.* 39, D670–D676.
- Schomburg, D., Schomburg, I., 2001–2006. *Springer Handbook of Enzymes*. 2nd ed. Springer, Heidelberg.
- Schomburg, D., Schomburg, I., BRENDA – from a database to a centre of excellence. *Systembiologie.de Int. Ed.* 10, 2016, 18–21.
- Schomburg, D., Salzmann, M., Stephan, D., 1990–1998. *Enzyme Handbook*, Vol. 1–17. Springer, Heidelberg.
- Schomburg, I., Hofmann, O., Baensch, C., Chang, A., Schomburg, D., 2000. Enzyme data and metabolic information: BRENDA, a resource for research in biology, biochemistry, and medicine. *Gene funct. Dis.* 3–4, 109–118.
- Schomburg, I., Chang, A., Hofmann, O., Ebeling, C., Ehrentreich, F., Schomburg, D., 2002. BRENDA: a resource for enzyme data and metabolic information. *Trends Biochem. Sci.* 27, 54–56.
- Schomburg, I., Chang, A., Placzek, S., Söhngen, C., Rother, M., Lang, M., Munaretto, C., Ulas, S., Stelzer, M., Grote, A., Scheer, M., Schomburg, D., 2013. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Res.* 41, D764–D772.
- Shin, S.-Y., Fauman, E.B., Petersen, A.-K., Krumsiek, J., Santos, R., Huang, J., Arnold, M., Erte, I., Forgetta, V., Yang, T.-P., Walter, K., Menni, C., Chen, L., Vasquez, L., Valdes, A.M., Hyde, C.L., Wang, V., Ziemeck, D., Roberts, P., Xi, L., Grundberg, E., Waldenberger, M., Richards, J.B., Mohney, R.P., Milburn, M.V., John, S.L., Trimmer, J., Theis, F.J., Overington, J.P., Suhre, K., Brosnan, M.J., Gieger, C., Kastenmüller, G., Spector, T.D., Soranzo, N., 2014. An atlas of genetic influences on human blood metabolites. *Nat. Genet.* 46, 543–550.
- Shlomi, T., Cabilio, M.N., Herrgard, M.J., Palsson, B.O., Ruppert, E., 2008. Network-based prediction of human tissue-specific metabolism. *Nat. Biotechnol.* 26, 1003–1010.
- Sillitoe, I., Lewis, T.E., Cuff, A., Das, S., Ashford, P., Dawson, N.J., Furnham, N., Laskowski, R.A., Lee, D., Lees, J.G., Lehtinen, S., Studer, R.A., Thornton, J., Orengo, C.A., 2015. CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res.* 43, D376–D381.
- Sumner, L.W., Mendes, P., Dixon, R.A., 2003. Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry* 62, 817–836.
- Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H., 2015. UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31 (31), 926–932.
- The Gene Ontology Consortium, 2015. Gene ontology consortium: going forward. *Nucleic Acids Res.* 43, D1049–D1056.
- The UniProt consortium, 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 45, D158–D169.
- Van Maris, A.J.A., Winkler, A.A., Kuypers, M., De Laat, W.T.A.M., Van Dijken, J.P., Pronk, J.T., 2007. Development of efficient xylose fermentation in *Saccharomyces cerevisiae*: xylose isomerase as a key component. *Adv. Biochem. Eng. Biotechnol.* 108, 179–204.
- Vary, P.S., Biedendieck, R., Fuerch, T., Meinhardt, F., Rohde, M., Deckwer, W.-D., Jahn, D., 2007. *Bacillus megaterium*-from simple soil bacterium to industrial protein production host. *Appl. Microbiol. Biotechnol.* 76, 957–967.

- Volkamer, A., Kuhn, D., Grombacher, T., Rippmann, F., Rarey, M., 2012. Combining global and local measures for structure-based druggability predictions. *J. Chem. Inf. Model.* 52, 360–372.
- Wittig, U., Kania, R., Golebiewski, M., Rey, M., Shi, L., Jong, L., Alga, E., Weidemann, A., Sauer-Danzwith, H., Mir, S., Krebs, O., Bittkowski, M., Wetsch, E., Rojas, I., Müller, W., 2012. SABIO-RK – database for biochemical reaction kinetics. *Nucleic Acids Res.* 40, D790–D796.
- Wittig, U., Rey, M., Kania, R., Bittkowski, M., Shi, L., Golebiewski, M., Weidemann, A., Müller, W., Rojas, I., 2014. Challenges for an enzymatic reaction kinetics database. *FEBS J.* 281, 572–582.
- Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., Kanehisa, M., 2008. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 24, i232–240.