Bioinformatics, 39(1), 2023, btac793
https://doi.org/10.1093/bioinformatics/btac793
Advance Access Publication Date: 9 December 2022
Original Paper

Databases and ontologies

Annotation of biologically relevant ligands in UniProtKB using ChEBI

Elisabeth Coudert [1], Sebastien Gehant [1], Edouard de Castro [1], Monica Pozzato[1], Delphine Baratin[1], Teresa Neto [1], Christian J. A. Sigrist [1], Nicole Redaschi [1], Alan Bridge [1,*]; and The UniProt Consortium

[1]Swiss-Prot Group, SIB Swiss Institute of Bioinformatics, Centre Medical Universitaire, 1211 Geneva 4, Switzerland, [2]European Molecular Biology Laboratory—European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridgeshire CB10 1SD, UK, [3]Protein Information Resource, University of Delaware, Newark, DE 19711, USA and [4]Protein Information Resource, Georgetown University Medical Center, Washington, DC 20007, USA

*To whom correspondence should be addressed.
Associate Editor: Peter Robinson

Received on August 19, 2022; revised on November 9, 2022; editorial decision on December 6, 2022; accepted on December 8, 2022

Abstract

Motivation: To provide high quality, computationally tractable annotation of binding sites for biologically relevant (cognate) ligands in UniProtKB using the chemical ontology ChEBI (Chemical Entities of Biological Interest), to better support efforts to study and predict functionally relevant interactions between protein sequences and structures and small molecule ligands.

Results: We structured the data model for cognate ligand binding site annotations in UniProtKB and performed a complete reannotation of all cognate ligand binding sites using stable unique identifiers from ChEBI, which we now use as the reference vocabulary for all such annotations. We developed improved search and query facilities for cognate ligands in the UniProt website, REST API and SPARQL endpoint that leverage the chemical structure data, nomenclature and classification that ChEBI provides.

Availability and implementation: Binding site annotations for cognate ligands described using ChEBI are available for UniProtKB protein sequence records in several formats (text, XML and RDF) and are freely available to query and download through the UniProt website (www.uniprot.org), REST API (www.uniprot.org/help/api), SPARQL endpoint (sparql.uniprot.org/) and FTP site (https://ftp.uniprot.org/pub/databases/uniprot/).

Contact: alan.bridge@sib.swiss

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The UniProt Knowledgebase (UniProtKB, at www.uniprot.org) is a reference resource of protein sequences and functional annotation that covers proteins from all branches of the tree of life (The UniProt Consortium, 2021). UniProtKB includes an expert-curated core of around 568 000 reviewed UniProtKB/Swiss-Prot protein sequence entries and over 229 million unreviewed UniProtKB/TrEMBL entries that are annotated by automatic systems (MacDougall *et al.*, 2020) (statistics for release 2022_04 of October 2022). UniProtKB provides a wealth of information on protein sequences and their functions, including the binding sites of biologically relevant or 'cognate' ligands (the term used in the remainder of this article) (Das and Orengo, 2018; Tyzack *et al.*, 2018) such as activators, inhibitors, cofactors and substrates, which are crucial to protein function. UniProt curators capture this knowledge through expert literature curation and from experimentally resolved protein structures in the protein data bank (PDB/PDBe) (Armstrong *et al.*, 2020; Burley *et al.*, 2021; Velankar *et al.*, 2021), removing adventitious ligands that are technical artefacts and mapping experimentally observed ligands in PDB to their cognate equivalents by reference to a curated list of known cognate ligands.

Here, we describe improvements to the annotation of cognate ligands and their binding sites in UniProtKB using the chemical ontology ChEBI (Chemical Entities of Biological Interest, www.ebi.ac.uk/chebi/) (Hastings *et al.*, 2016). We have performed a complete reannotation of cognate ligands and their binding sites in UniProtKB, replacing textual descriptions of ligands with stable unique identifiers from the ChEBI ontology, and now use ChEBI as the reference vocabulary for all new cognate ligand binding site annotations. This work makes knowledge of cognate ligands and

their binding sites in UniProtKB easier to find and access. It provides improved support for the design of biochemical experiments (Fleischhacker *et al.*, 2015; Frederick *et al.*, 2022) and computational approaches (Das *et al.*, 2021; Littmann *et al.*, 2021; Wehrspan *et al.*, 2022; Wu *et al.*, 2018) to elucidate protein functions and interactions, and enhances interoperability with other resources providing knowledge of cognate ligands such as PDBe (Mukhopadhyay *et al.*, 2019), BioLiP (Yang *et al.*, 2013), FireDB (Maietta *et al.*, 2014), MetalPDB (Putignano *et al.*, 2018) and PDBBind (Liu *et al.*, 2015).

2 Materials and methods

2.1 Changes to the UniProt data model and formats

Most sequence annotations (also called 'features') in UniProtKB, including the cognate ligand binding site annotations that are the subject of this work, consist of three main elements. The 'feature location' defines the sequence region or amino acid residue position that is annotated, the 'feature key' specifies the type of each feature, and the 'feature description' provides a textual description, which for cognate ligands includes the name of the ligand and other relevant information, such as numbering (of multiple ligands of the same type) and ligand roles. We structured this description for binding site annotations into several fields (described in the online documentation at www.uniprot.org/release-notes/2022-08-03-release), to standardize the description of a ligand, and optionally the bound part of the ligand (such as the iron atom in a heme, or a nucleotide in a macromolecule such as DNA), with the ChEBI ontology. We illustrate this new data model with examples in Section 3. We also simplified the range of feature keys that are used for cognate ligand binding site annotations, which prior to this work were the following:

- 'CA_BIND', which denotes a sequence region that binds to calcium;
- 'METAL', which denotes a sequence position that binds a metal;
- 'NP_BIND', which denotes a sequence region that binds a nucleotide phosphate;
- 'BINDING', which denotes a sequence position that binds any type of chemical entity;
- 'REGION', which denotes a sequence region of interest in a protein (including a region that binds a ligand).

The ChEBI ontology provides a means to search for any ligand or class of ligand represented in ChEBI, at any desired level of specificity, without requiring ligand-specific feature keys. We therefore deprecated the feature keys 'CA_BIND', 'METAL' and 'NP_BIND', and now use the feature key 'BINDING' for all binding site annotations for all cognate ligands. We also recurated all cognate ligand binding sites of interest found in features of the type 'REGION' and moved them to 'BINDING'. Finally, we modified all UniRules (MacDougall *et al.*, 2020), including HAMAP (Pedruzzi *et al.*, 2015) and PROSITE (Sigrist *et al.*, 2013) rules, to provide binding site annotations using ChEBI identifiers in the new data model described here.

2.2 Mapping of legacy text annotations of cognate ligand binding sites in UniProtKB to ChEBI

UniProtKB previously described cognate ligands in binding site annotations using text labels, such as 'lipid', 'cholesterol', 'heme', 'heme b', 'divalent metal' or 'zinc'. To standardize the descriptions of biologically relevant ligands in binding site annotations in UniProtKB, we created a one-to-one mapping between each such text label and the corresponding ChEBI identifier and used that mapping to reannotate all legacy data.

We extracted unique ligand descriptions from binding site annotations linked to each of the feature keys 'CA_BIND', 'METAL', 'NP_BIND' and 'BINDING', as well as 'REGION' annotations with the word 'binding' in the feature description, and mapped each of the text labels found to the corresponding ChEBI identifier manually. During the mapping, we selected the ChEBI that represents the major microspecies of the ligand (the predominant protonation state) at pH 7.3, which is the convention used in UniProtKB and the Rhea reaction knowledgebase (www.rhea-db.org) (Bansal *et al.*, 2022). If an appropriate ChEBI entity was not already available, then we submitted the required structure to ChEBI for inclusion in the chemical ontology. We also assigned a 'UniProt name' to each ChEBI entity used in our annotations, which as its name suggests, is a specific synonym that is created and used by UniProt (and is also used in Rhea).

Some ligand text labels presented with multiple possible mappings to ChEBI—sometimes due to stereochemistry issues—while some described generic classes of chemicals or roles such as cofactor, hormone or odorant, which we could not map to any defined structure. We examined each of these cases in turn and, where necessary, recurated them, using information from the literature, the PDB and the UniProtKB protein sequence records concerned, including existing Rhea reaction annotations, before selecting the most appropriate mapping to ChEBI. In total, we recurated over 100 such ambiguous ligands.

Once complete, we used the mapping of defined cognate ligands to replace legacy text labels in UniProtKB with the corresponding identifiers from ChEBI. We also used additional information from the existing annotations, such as ligand numbering and roles, to populate the corresponding data fields in the new structured data model.

We did not yet systematically recurate binding site annotations for enzymes in UniProtKB with the generic text label 'substrate', which does not specify which of the possible substrate(s) are bound. We are continuing to map these legacy 'substrate' annotations to specific ChEBI identifiers, using Rhea annotations and other information such as ligand data from PDBe records where available, mapped to UniProt sequences using the SIFTS framework (Dana *et al.*, 2019).

2.3 UniProt tools and services to exploit ligand binding site annotations

We modified the UniProt website www.uniprot.org, UniProt REST API www.uniprot.org/help/api and UniProt SPARQL endpoint sparql.uniprot.org/, to support searches for ligand binding site annotations using ChEBI identifiers, ligand names, synonyms and chemical structures from ChEBI encoded as InChIKeys. The InChIKey is a simple hash representation of chemical structures that provides a convenient means to search and map chemical structure databases (see www.inchi-trust.org/).

3 Results

3.1 Structuring cognate ligand binding site annotations in UniProtKB using ChEBI

The annotation of cognate ligand binding sites in UniProtKB using the chemical ontology ChEBI was made available from UniProt release 2022_03 of August 2022. This initial release featured 776 unique ligands from ChEBI, which were involved in over 980 000 binding site annotations for over 200 000 UniProtKB/Swiss-Prot protein sequence records, and over 65 million binding site annotations for over 17 million protein sequence records for the whole of UniProtKB, including UniProtKB/TrEMBL. We provide a complete list of all cognate ligands used in binding site annotations in UniProtKB release 2022_03 in Supplementary Table S1. This list is part of a larger set of allowed ligands for binding site annotations in UniProtKB, which also includes all ChEBI entities used in Rhea reactions.

The new data model improves the consistency of annotations while retaining flexibility. It supports the annotation of binding sites for ligands described at any level of granularity in ChEBI, from broad classes of ligands such as 'metal cation' (CHEBI: 25213) or 'heme' (CHEBI: 30413), to structurally defined ligands such as

'Fe(2+)' (CHEBI: 29033) or 'heme b' (CHEBI: 60344). It also supports the annotation of binding sites for ligands that are parts of larger macromolecules. The example below shows one such case, where amino acid 146 of yeast L-lactate dehydrogenase (UniProtKB/Swiss-Prot entry P00175) binds to the iron atom (CHEBI: 18248) of heme b (CHEBI: 60344) (this form of heme b represents the predominant protonation state at pH7.3, the form chosen by convention in UniProtKB). The 'evidence' field lists the evidences that support the annotation. Each evidence is described by a term from the Evidence and Conclusions Ontology ECO (Nadendla *et al.*, 2022), and the source of the information, here experiments published in two peer-reviewed articles (Cunane *et al.*, 2002; Xia and Mathews, 1990) and protein structures 1FCB and 1KBI from the PDB.

```
FT   BINDING         146
FT                   /ligand='heme b'
FT                   /ligand_id='ChEBI:CHEBI:60344'
FT                   /ligand_part='Fe'
FT                   /ligand_part_id='ChEBI:CHEBI:18248'
FT                   /note='axial binding residue'
FT                   /evidence='ECO:0000269|PubMed:11914072,
FT                   ECO:0000269|PubMed:2329585,
FT                   ECO:0007744|PDB:1FCB,
FT                   ECO:0007744|PDB:1KBI'
```

We refer readers to the online documentation at www.uniprot.org/release-notes/2022-08-03-release, which provides additional examples of binding site annotations in the UniProtKB formats text, XML and RDF/XML.

3.2 UniProt tools and services to access and query cognate ligand binding site annotations made with ChEBI

Users can access and query UniProtKB cognate ligand binding site annotations made with ChEBI using the UniProt website, REST API and SPARQL endpoint.

3.2.1 UniProt website

The UniProt website www.uniprot.org provides access to UniProtKB protein sequence records and annotations, including cognate ligand binding site annotations for each protein (Fig. 1). Users can now query the website for proteins that bind cognate ligands of interest using identifiers, names, synonyms and chemical structures (encoded as InChIKeys) from ChEBI using the advanced query builder. The complete ChEBI ontology is indexed, so that searches using identifiers for higher-level grouping classes in the ChEBI ontology will retrieve UniProtKB records with binding site annotations to all child classes. ChEBI identifiers entered by users are automatically mapped to those of the major microspecies at pH 7.3, which is the form used in UniProtKB and Rhea, using a mapping file provided by Rhea.

The sample query shown below will retrieve all proteins with binding site annotations for any kind of heme, using the ChEBI identifier for that grouping class, which is ChEBI: 30413:

1. (ft_binding:'CHEBI: 30413')
   The result set will include all proteins with binding site annotations for any form of heme (CHEBI: 30413), including 'heme b' (CHEBI: 60344), 'heme c' (CHEBI: 61717) and all others. To retrieve proteins with binding site annotations for specific forms of heme, users can simply change the ChEBI identifier to that of the form desired, here ChEBI: 60344 for 'heme b':
2. (ft_binding: 'CHEBI: 60344')
   Users can also perform searches for specific ligands using the chemical structure represented as an InChIKey, if using a chemical structure database other than ChEBI:
3. (ft_binding: KABFMIBPWCXCRK-RGGAHWMASA-J)
   Users can elect to ignore the charge, by removing the third block of the InChIKey, as in this example:
4. (ft_binding: KABFMIBPWCXCRK-RGGAHWMASA)
   They may also elect to ignore both stereochemistry and charge, by removing both the second and third blocks of the InChIKey:
5. (ft_binding: KABFMIBPWCXCRK)
   Users can also combine searches for cognate ligand binding site annotations with other types of annotations, as in this query for human mitochondrial flavoproteins (i.e. proteins with annotated binding sites for some CHEBI: 30527 - flavin) that are linked to genetic diseases defined by the resource Online Mendelian Inheritance in Man (Hamosh *et al.*, 2021):
6. (ft_binding: 'CHEBI: 30527') AND (cc_scl_term: SL-0173) AND (organism_id: 9606) AND (cc_disease:*)

We provide complete documentation on searching for small molecule data in UniProtKB, including ligands described in binding site annotations, at www.uniprot.org/help/chemical_data_search.

3.2.2 UniProt REST API

The UniProt REST API (www.uniprot.org/help/api) allows users to query and process UniProt data programmatically and to specify the required output format for query results (such as txt, xml, rdf, tsv, etc.) and, for the tab-separated format, the desired annotation fields. The simplest way to create URLs for programmatic use is by using the advanced query builder to set the desired query fields and values, perform the search and click the 'Download' button, which opens a panel with a 'Generate URL for API' link. Users can now query the UniProt REST API with identifiers, names, synonyms and chemical structures from ChEBI for ligand-binding site annotations.

3.2.3 UniProt SPARQL endpoint

The UniProt SPARQL endpoint sparql.uniprot.org allows users to query UniProt RDF data and RDF data from other SPARQL endpoints using federated SPARQL queries. It now supports queries for ligand-binding site annotations using identifiers, names, synonyms and chemical structure data from ChEBI. We demonstrate this capability using a federated SPARQL query that combines the UniProt SPARQL endpoint and that of the Integrated Database of Small Molecules (IDSM) (Galgonek and Vondrášek, 2021; Kratochvil *et al.*, 2019). IDSM supports fingerprint-guided chemical similarity and substructure searches in a number of chemical datasets, including ChEBI, using Sachem, a high-performance open source chemical cartridge (Kratochvil *et al.*, 2018). This federated SPARQL query allows UniProt to borrow that functionality from IDSM; it will find all proteins that bind to ligands with structures similar to that of a query ligand, in this case, heme b (specified using SMILES or Simplified Molecular-Input Line-Entry notation) (http://opensmiles.org). The UniProt SPARQL endpoint queries that of IDSM, which returns the set of chemical entities in ChEBI that are similar to the query ligand heme b (above a Tanimoto similarity score threshold of 0.8) and then searches for proteins in UniProtKB with binding site annotations for those ligands, which it then returns to the user.

```
SELECT? uniprot? mnemonic? proteinName? ligandSimilarityScore?
ligand
WHERE {
  SERVICE <https://idsm.elixir-czech.cz/sparql/endpoint/chebi> {
    [sachem: compound? ligand; sachem: score? ligandSimilarityScore]
    sachem: similaritySearch
      [
      sachem: query
          'CC1 = C(CCC([O-])=O)C2=[N+]3C1 = Cc1c(C)c(C =
           C)c4C=C5C(C)=C(C = C)C6=[N+]5[Fe-
          ]3(n14)n1c(=C6)c(C)c(CCC([O-])=O)c1 = C2';
      sachem: cutoff '8e-1'^^ xsd: double;
      sachem: aromaticityMode sachem: aromaticityDetect;
```

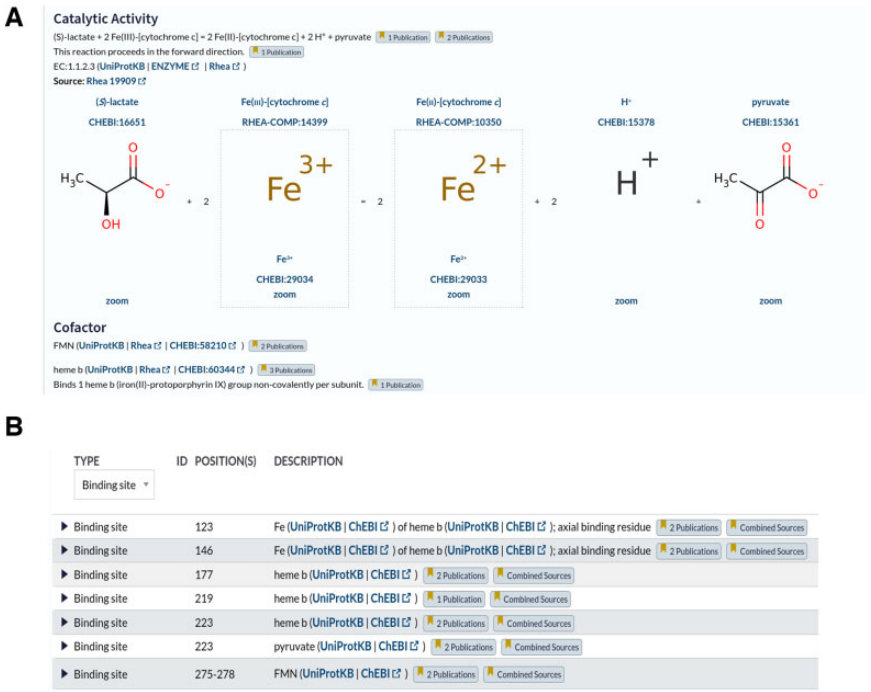

Fig. 1. Website view of small molecule annotations in UniProtKB, including cognate ligand binding site annotations—from www.uniprot.org/uniprotkb/P00175/entry. All small molecule annotations are shown in the 'Function' section. (**A**) The 'Catalytic Activity' subsection describes enzymatic reactions using Rhea (which is based on ChEBI), while cofactors are described using ChEBI in the 'Cofactor' subsection. Standardization of reaction and cofactor descriptions was performed in previous work, and is shown here for completeness. (**B**) The 'Features' subsection displays the available binding site annotations for cognate ligands described using ChEBI, the subject of this work. Each ligand has a link 'UniProtKB' to launch searches for other proteins binding this ligand, a link out to 'ChEBI', and expandable sections like 'Publications' to examine provenance and evidence

```
        sachem: similarityRadius 1;
        sachem: tautomerMode sachem: ignoreTautomers
      ]
    }
    ? uniprot up: mnemonic? mnemonic.
    ? uniprot up: recommendedName/up: fullName? proteinName.
    ? uniprot up: annotation? annotation.
    ? annotation a up: Binding_Site_Annotation;
    up: ligand/rdfs: subClassOf? ligand.
}
ORDER BY DESC(? ligandSimilarityScore)
```

This type of query could be useful in the study of 3D protein structures and protein structure models. Given the SMILES representation of a non-cognate ligand from an experimentally determined 3D protein structure from PDBe, users can retrieve similar cognate ligands from UniProtKB that could replace it to create a more biologically relevant structure. Predicted 3D protein structure models from state-of-the-art methods such as AlphaFold (Jumper *et al.*, 2021; Varadi *et al.*, 2022) lack ligands altogether, and methods that transfer experimental ligands from similar structures in PDBe (Hekkelman *et al.*, 2022) might exploit UniProtKB as a source of cognate ligands for this transfer. We provide more sample queries in the online documentation for the UniProt SPARQL endpoint at https://sparql.uniprot.org/.well-known/sparql-examples/, while the developers of the IDSM SPARQL endpoint provide additional documentation at https://idsm.elixir-czech.cz/sparql/doc/manual.html.

4 Conclusions and future work

We have structured and reannotated cognate ligand binding sites in UniProtKB using ChEBI and report new tools and services to exploit this improved ligand dataset via the UniProt website and APIs. This work is part of an ongoing program to standardize all small molecule annotations in UniProtKB using ChEBI and builds on previous improvements to the annotation of enzymes and transporters in UniProtKB using the Rhea knowledgebase of biochemical reactions, which uses ChEBI to represent reactants (Bansal *et al.*, 2022; Morgat *et al.*, 2020). We continue to work to improve the UniProtKB cognate ligand dataset through expert literature curation, supported by machine learning approaches to target relevant literature for information extraction (Allot *et al.*, 2021; Islamaj *et al.*, 2021), and through the development of improved pipelines for the import and curation of ligand data from PDBe.

Acknowledgements

We thank the Cheminformatics and Metabolism Team of EMBL-EBI for their work in maintaining and developing ChEBI, particularly Adnan Malik and Gareth Owen for expert curation and advice and Andrew Leach, and the PDBe team at EMBL-EBI, for maintaining and developing PDBe, the principal source of ligand data for UniProtKB. We also thank Sameer Velankar of PDBe at EMBL-EBI for stimulating discussions on methods to identify cognate ligands in protein structures in PDBe. We gratefully acknowledge the software contributions of ChemAxon (https://www.chemaxon.com/products/marvin/).

Funding

UniProt was supported by the National Eye Institute (NEI), National Human Genome Research Institute (NHGRI), National Heart, Lung, and Blood Institute (NHLBI), National Institute on Aging (NIA), National Institute of Allergy and Infectious Diseases (NIAID), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of General Medical Sciences (NIGMS), National Institute of Mental Health (NIMH) and National Cancer Institute (NCI), and Office of the Director of the National Institutes of Health (NIH) [U24HG007822]; National Human Genome Research Institute [U41HG002273]; National Institute of General Medical

Sciences [R01GM080646 and P20GM103446]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. UniProt activities at the SIB were also supported by the Swiss Federal Government through the State Secretariat for Education, Research and Innovation SERI. Additional support for the EMBL-EBI's involvement in UniProt comes from European Molecular Biology Laboratory (EMBL) core funds, the Alzheimer's Research UK (ARUK) [ARUK-NAS2017A-1], the Biotechnology and Biological Sciences Research Council (BBSRC) [BB/T010541/1] and Open Targets. PIR's UniProt activities were also supported by the NIH grants [R01GM080646, G08LM010720 and P20GM103446], and the National Science Foundation (NSF) grant [DBI-1062520].

Conflict of Interest: none declared.

References

Allot,A. *et al.* (2021) LitSuggest: a web-based system for literature recommendation and curation using machine learning. *Nucleic Acids Res.*, **49**, W352–W358.

Armstrong,D.R. *et al.* (2020) PDBe: improved findability of macromolecular structure data in the PDB. *Nucleic Acids Res.*, **48**, D335–D343.

Bansal,P. *et al.* (2022) Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res.*, **50**, D693–D700.

Burley,S.K. *et al.* (2021) RCSB protein data bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.*, **49**, D437–D451.

Cunane,L.M. *et al.* (2002) Crystallographic study of the recombinant flavin-binding domain of baker's yeast flavocytochrome b(2): comparison with the intact wild-type enzyme. *Biochemistry*, **41**, 4264–4272.

Dana,J.M. *et al.* (2019) SIFTS: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res*, **47**, D482–D489.

Das,S. and Orengo,C. (2018) Choosing the best enzyme complex structure made easy. *Structure*, **26**, 528–530.

Das,S. *et al.* (2021) CATH functional families predict functional sites in proteins. *Bioinformatics*, **37**, 1099–1106.

Fleischhacker,A.S. *et al.* (2015) The C-terminal heme regulatory motifs of heme oxygenase-2 are redox-regulated heme binding sites. *Biochemistry*, **54**, 2709–2718.

Frederick,A.K. *et al.* (2022) Effect on intrinsic peroxidase activity of substituting coevolved residues from omega-loop C of human cytochrome c into yeast iso-1-cytochrome c. *J. Inorg. Biochem.*, **232**, 111819.

Galgonek,J. and Vondrášek,J. (2021) IDSM ChemWebRDF: SPARQLing small-molecule datasets. *J. Cheminform.*, **13**, 38.

Hamosh,A. *et al.* (2021) Online Mendelian Inheritance in Man (OMIM(R)): Victor McKusick's magnum opus. *Am. J. Med. Genet. A*, **185**, 3259–3265.

Hastings,J. *et al.* (2016) ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res.*, **44**, D1214–1219.

Hekkelman,M.L. *et al.* (2022) AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat. Methods*,

Islamaj,R. *et al.* (2021) NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. *Sci. Data*, **8**, 91.

Jumper,J. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.

Kratochvil,M. *et al.* (2018) Sachem: a chemical cartridge for high-performance substructure search. *J. Cheminform.*, **10**, 27.

Kratochvil,M. *et al.* (2019) Interoperable chemical structure search service. *J. Cheminform.*, **11**, 45.

Littmann,M. *et al.* (2021) Protein embeddings and deep learning predict binding residues for various ligand classes. *Sci. Rep.*, **11**, 23916.

Liu,Z. *et al.* (2015) PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, **31**, 405–412.

MacDougall,A. UniProt Consortium. *et al.* (2020) UniRule: a unified rule resource for automatic annotation in the UniProt knowledgebase. *Bioinformatics*, **36**, 4643–4648.

Maietta,P. *et al.* (2014) FireDB: a compendium of biological and pharmacologically relevant ligands. *Nucleic Acids Res.*, **42**, D267–D272.

Morgat,A. *et al.*; UniProt Consortium. (2020) Enzyme annotation in UniProtKB using rhea. *Bioinformatics*, **36**, 1896–1901.

Mukhopadhyay,A. *et al.* (2019) Finding enzyme cofactors in protein data bank. *Bioinformatics*, **35**, 3510–3511.

Nadendla,S. *et al.* (2022) ECO: the evidence and conclusion ontology, an update for 2022. *Nucleic Acids Res.*, **50**, D1515–D1521.

Pedruzzi,I. *et al.* (2015) HAMAP in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Res.*, **43**, D1064–D1070.

Putignano,V. *et al.* (2018) MetalPDB in 2018: a database of metal sites in biological macromolecular structures. *Nucleic Acids Res.*, **46**, D459–D464.

Sigrist,C.J. *et al.* (2013) New and continuing developments at PROSITE. *Nucleic Acids Res.*, **41**, D344–347.

The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*, **49**, D480–D489.

Tyzack,J.D. *et al.* (2018) Ranking enzyme structures in the PDB by bound ligand similarity to biological substrates. *Structure*, **26**, 565–571 e563.

Varadi,M. *et al.* (2022) AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.

Velankar,S. *et al.* (2021) The protein data bank archive. *Methods Mol. Biol.*, **2305**, 3–21.

Wehrspan,Z.J. *et al.* (2022) Identification of iron-sulfur (Fe-S) cluster and zinc (Zn) binding sites within proteomes predicted by DeepMind's AlphaFold2 program dramatically expands the metalloproteome. *J. Mol. Biol.*, **434**, 167377.

Wu,Q. *et al.* (2018) COACH-D: improved protein-ligand binding sites prediction with refined ligand-binding poses through molecular docking. *Nucleic Acids Res.*, **46**, W438–W442.

Xia,Z.X. and Mathews,F.S. (1990) Molecular structure of flavocytochrome b2 at 2.4 a resolution. *J. Mol. Biol.*, **212**, 837–863.

Yang,J. *et al.* (2013) BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.*, **41**, D1096–1103.