ModelCIF: An extension of PDBx/mmCIF data representation for computed structure models

SUPPLEMENTARY MATERIAL

Methods

Extending PDBx/mmCIF for computed structure models

ModelCIF is an extension of PDBx/mmCIF [1, 2] for describing macromolecular structures modeled by computational methods. The definitions in ModelCIF are the result of collaborations and discussions among the ModelCIF working group members (wwpdb.org/task/modelcif) and are based on examples of computed structure models (CSMs) obtained from modeling resources such as SWISS-MODEL repository [3], RoseTTAFold [4, 5], ModelArchive (https://www.modelarchive.org), ModBase [6], I-TASSER [7, 8], and AlphaFoldDB [9]. These collaborative efforts have culminated in the creation of version 1.4.4 of the ModelCIF dictionary extension released in December 2022 (https://mmcif.wwpdb.org/dictionaries/mmcif_ma.dic/Index/). ModelCIF is licensed under Creative Commons Zero v1.0 Universal (CC0 1.0) license.

ModelCIF dictionary development (GitHub)

The ModelCIF dictionary is developed as a collaborative project on a public GitHub repository (https://github.com/ihmwg/ModelCIF). The complete ModelCIF dictionary extension file is available under the distribution (i.e., "dist") directory (mmcif_ma.dic) within the repository. This complete and distributable form of the extension is constructed from the concatenation of the individual source dictionary components found under the "base" directory (https://github.com/ihmwg/ModelCIF/tree/master/base); thus, these "base" files are what are directly modified by maintainers, which are then combined to generate the complete distributable extension file (using the "build" scripts in https://github.com/ihmwg/ModelCIF/tree/master/scripts). Any additions or modifications arising from the discussions in the ModelCIF WG are incorporated in the source dictionary files on GitHub, followed by a full update and release of the dictionary. Whenever ModelCIF is updated, the latest PDBx/mmCIF dictionary is pulled from the corresponding Git repository (https://github.com/wwpdb-dictionaries/mmcif_pdbx) to ensure compatibility and compliance with the latest PDBx/mmCIF definitions. Importantly, upon each update, the ModelCIF version number is incremented and metadata regarding the revision history is included.

ModelCIF validation and maintenance (GitHub, mmCIF resources website, PDBx/mmCIF dictionary tools)

Following the introduction of any new modifications and version updates to ModelCIF, the resulting fully-formed dictionary undergoes a series of quality checks to ensure the dictionary is formatted properly, is free of any syntax errors, and successfully validates a set of example model files. Some of these checks are performed while building the concatenated dictionary file from the source "base" files described above. In addition, PDBx/mmCIF software tools (e.g., https://github.com/rcsb/cpp-dict-pack) are used to further validate the updated dictionary. Only

upon successful passing of these validation tests is the new version of the dictionary extension released on GitHub and published on the PDBx/mmCIF resources website, mmcif.wwpdb.org (i.e., https://mmcif.wwpdb.org/dictionaries/mmcif_ma.dic/Index/).

Results and discussion

Supporting software tools and resources

The following software tools and CSM resources support ModelCIF (also listed in Table 1 in the main text).

python-modelcif: The python-modelcif library (https://github.com/ihmwg/python-modelcif) provides a mechanism to describe a structural model as a hierarchy of Python objects. These objects capture the same information as in the ModelCIF dictionary and can be written out to or read in from either mmCIF or BinaryCIF [10] files (BinaryCIF format provides improved compression and parsing performance). The library automatically handles the various object identifiers used in the ModelCIF dictionary, mapping them naturally to interrelationships between Python objects, and mapping controlled vocabulary terms defined in the dictionary to Python subclasses. The library thus generates conformant ModelCIF files "by construction", although it is also able to read in and validate files against the ModelCIF dictionary. Convenience Python application programming interfaces (APIs) are also provided to add additional metadata to output files, such as citation information from PubMed, reference sequence information from UniProtKB [11], and metadata about associated files. The library is designed such that it can be embedded in a modeling or visualization package to support input or output of ModelCIF files, or it can be used standalone, and it is available under the permissive MIT open-source license. The python-modelcif library provides scripts to (a) convert ModelCIF files to BinaryCIF format (https://raw.githubusercontent.com/ihmwg/python-modelcif/main/examples/convert_bcif.py); (b) validate a ModelCIF file against the dictionary (https://raw.githubusercontent.com/ihmwg/python-modelcif/main/examples/validate_mmcif.py); and (c) create a minimal ModelCIF file without any additional metadata (https://raw.githubusercontent.com/ihmwg/python-modelcif/main/util/make-mmcif.py).

ModelCIF-converters and other tools that support ModelCIF: ModelCIF-converters (https://git.scicore.unibas.ch/schwede/modelcif-converters) is a growing repository of scripts and use cases for converting models from legacy PDB format to ModelCIF using python-modelcif and OpenStructure [12]. The repository also includes a containerized solution for validating ModelCIF files. In addition, there are other tools that convert PDB format files generated from specific modeling applications to ModelCIF (https://github.com/salilab/modbase_utils, https://github.com/ihmwg/modelcif_utils).

wwPDB mmCIF software resources webpage: This website provides a list of software libraries and tools that support PDBx/mmCIF (https://mmcif.wwpdb.org/docs/software-resources.html), many of which also support ModelCIF (e.g., ciftools-java, py-mmcif). These include tools developed by the wwPDB and the structural and computational biology community.

ModelArchive: ModelArchive supports the submission of ModelCIF files. The information necessary for deposition is extracted from the metadata in the ModelCIF file. ModelCIF files are also mandatory for the deposition of large datasets. Current datasets include structural predictions of core eukaryotic protein complexes generated by RoseTTAFold [4, 5] in combination with AlphaFold2 [13] (ma-bak-cepc) [5], structural and functional predictions using ColabFold in combination with FoldSeek (ma-coffe-slac), structural predictions for *Mycobacterium tuberculosis* proteins of relevance for antibiotic resistance obtained using a combination of SWISS-MODEL [14], AlphaFold-Multimer and small molecule docking (ma-tbvar3d), and AlphaFold2 predictions of *Sphagnum divinum* (ma-ornl-sphdiv) [15] and the African swine fever virus (ma-asfv-asfvg). To assist depositions and to generate correctly formatted ModelCIF files, the ModelArchive team developed ModelCIF-converters (https://git.scicore.unibas.ch/schwede/modelcif-converters). ModelArchive furthermore takes advantage of cross-references to UniProtKB [11] included in ModelCIF files to connect these models within the 3D-Beacons network [16], which provides unified data access to models from various model providers.

SWISS-MODEL: The ModelCIF format is well suited to provide a single output file for homology models generated by SWISS-MODEL [14]. Experimental support for ModelCIF files is available for interactive SWISS-MODEL projects and includes information on templates and alignments used as well as model quality assessments. Work is in progress to add ModelCIF support to the SWISS-MODEL Repository [3].

Modeller: The Modeller [17] comparative modeling pipeline supports output of models in ModelCIF compliant format, including information on the template(s) and alignment(s) used.

Zhang-Group servers: Most protein structure prediction servers from Zhang-Group now support ModelCIF. For example, I-TASSER related servers [7, 8], QUARK related servers [18, 19], and the LOMETS [20] server can provide the final predicted model in both PDB format [21] and ModelCIF format. In addition to the model coordinates information, the ModelCIF-formatted output also includes multiple sequence alignments, structural templates, deep learning predicted spatial restraints, and both global and per-residue model quality assessment scores.

AlphaFoldDB: AlphaFoldDB [9] is a database of protein structures predicted using AlphaFold2 [13] and as of October 2022 it contains predicted monomer structures for over 214 million UniProtKB [11] sequences. AlphaFoldDB uses ModelCIF as the primary data format. It is used for visualization of the predicted structures on the website and distribution of the data to users. Since the July 2022 release, all AlphaFoldDB entries are ModelCIF compliant. AlphaFoldDB ModelCIF files provide the 3D coordinate data together with the predicted LDDT (pLDDT) scores [22], secondary structure information, a list of template structures provided as extra input to the AlphaFold2 model during structure prediction, and metadata about the sequence extracted from UniProtKB. The Predicted Aligned Error (PAE) is currently provided in a separate JavaScript Object Notation (JSON) file. AlphaFoldDB uses the py-mmcif library (https://github.com/rcsb/py-mmcif) to convert ModelCIF files to BinaryCIF format for improving storage size and parsing performance. A subset of the data available in the ModelCIF file is also provided in the PDB format

for every AlphaFoldDB entry, but only for compatibility with legacy tools. New data and metadata will be added only in the ModelCIF files going forward. AlphaFoldDB models are available via the 3D-Beacons network [16], which provides unified data access to models from various model providers.

RoseTTAFold: RoseTTAFold [4, 5] is a three-track neural network model which enables rapid prediction of atomic structures for single protein chains as well as protein-protein complexes. RoseTTAFold is the default modeling protocol in the Robetta protein structure prediction server (https://robetta.bakerlab.org/) [23], but it can also be deployed locally (https://github.com/RosettaCommons/RoseTTAFold). At the time of publication, RoseTTAFold module of the Robetta server will support ModelCIF.

Structure visualization tools: CSMs can be visualized using structure visualization tools such as Mol* [24] and ChimeraX [25]. Both Mol* and ChimeraX support ModelCIF. In addition to visualizing atomic structures, local model quality score-based coloring is supported. ChimeraX also supports visualization of target-template alignments described in ModelCIF.

Examples of data and metadata represented in ModelCIF

Targets, templates, alignments, and external references: Adapted from ModelArchive entry
ma-tbvar3d-18

```
_ma_target_ref_db_details.target_entity_id                  1
_ma_target_ref_db_details.db_name                           UNP
_ma_target_ref_db_details.db_code                           CLPC1_MYCTU
_ma_target_ref_db_details.db_accession                      P9WPC9
_ma_target_ref_db_details.seq_db_align_begin                1
_ma_target_ref_db_details.seq_db_align_end                  848
_ma_target_ref_db_details.ncbi_taxonomy_id                  83332
_ma_target_ref_db_details.organism_scientific               'Mycobacterium
tuberculosis (strain ATCC 25618 / H37Rv)'
_ma_target_ref_db_details.seq_db_sequence_version_date  2014-04-16
_ma_target_ref_db_details.seq_db_sequence_checksum      4D536FBDA183CD94


loop_
_ma_data.id
_ma_data.name
_ma_data.content_type
1 'Template chain 5OG1, label_asym_id=A' 'template structure'
2 'CLPC1_MYCTU (P9WPC9)' target
3 'Target-template alignment by BLAST' 'target-template alignment'
4 'Model for P9WPC9' 'model coordinates' .

_ma_template_details.ordinal_id                     1
_ma_template_details.template_id                    1
_ma_template_details.template_origin                'reference database'
_ma_template_details.template_entity_type           polymer
_ma_template_details.template_data_id               1
_ma_template_details.target_asym_id                 A
_ma_template_details.template_label_asym_id         A
_ma_template_details.template_label_entity_id       1
_ma_template_details.template_model_num             1
_ma_template_details.template_auth_asym_id          C

_ma_template_poly_segment.id                        1
_ma_template_poly_segment.template_id               1
_ma_template_poly_segment.residue_number_begin      7
_ma_template_poly_segment.residue_number_end        861

_ma_target_template_poly_mapping.id                     1
_ma_target_template_poly_mapping.template_segment_id    1
_ma_target_template_poly_mapping.target_asym_id         A
_ma_target_template_poly_mapping.target_seq_id_begin    1
_ma_target_template_poly_mapping.target_seq_id_end      848
```

```
_ma_template_ref_db_details.template_id            1
_ma_template_ref_db_details.db_name                PDB
_ma_template_ref_db_details.db_accession_code      5OG1
_ma_template_ref_db_details.db_version_date        2019-08-21


_ma_alignment_info.alignment_id      1
_ma_alignment_info.data_id           3
_ma_alignment_info.alignment_length 917
_ma_alignment_info.alignment_type    'target-template pairwise alignment'
_ma_alignment_info.alignment_mode    local


_ma_alignment_details.ordinal_id                 1
_ma_alignment_details.alignment_id               1
_ma_alignment_details.template_segment_id        1
_ma_alignment_details.target_asym_id             A
_ma_alignment_details.score_type                 'BLAST e-value'
_ma_alignment_details.score_value                0
_ma_alignment_details.percent_sequence_identity 46.784


loop_
_ma_alignment.ordinal_id
_ma_alignment.alignment_id
_ma_alignment.target_template_flag
_ma_alignment.sequence
1 1 1
;MFERFTDRARRVVVLAQEEARMLNHNYIGTEHILLGLIHEGEGVAAKSLESLGISLEGVRSQVEEIIGQGQQA--P
SGHIPFTPRAKKVLELSLREALQLGHNYIGTEHILLGLIREGEGVAAQVLVKLGAELTRVRQQVIQLLSGYQGKEAA
EAGTGGRGGESGSPSTSLVLDQFGRNLTAAAMEGKLDPVIGREKEIERVMQVLSRRTKNNPVLIGEPGVGKTAVVEG
LAQAIVHGEVPETLKDKQLYTLDLGSLVAGSRYRGDFEERLKKVLKEINTR-GDIILFIDELHTLVGAGAAEGAIDA
ASILKPKLARGELQTIGATTLDEYRKYIEKDAALERRFQPVQVGEPTVEHTIEILKGLRDRYEAHHRVSITDAAMVA
AATLADRYINDRFLPDKAIDLIDEAGARMRIRRMTAPPDLREFDEKIAEARREKESAIDAQDFEKAA-------SLR
DREKTLVAQRAEREKQW-------------------------RSGDLDVVAE--------------------
--------VDDEQIAEVLGNWTGIPVFKLTEAETTRLLRMEEELHKRIIGQEDAVKAVSKAIRRTRAGLKDPKRPSG
SFIFAGPSGVGKTELSKALANFLFGDDDALIQIDMGEFHDRFTASRLFGAPPGYVGYEEGGQLTEKVRRKPFSVVLF
DEIEKAHQEIYNSLLQVLEDGRLTDGQGRTVDFKNTVLIFTSNLGTSDIS-KPVGLGFSKGGGENDYERMKQKVNDE
LKKHFRPEFLNRIDDIIVFHQLTREEIIRMVDLMISRVAGQLKSKDMALVLTDAAKALLAKRGFDPVLGARPLRRTI
QREIEDQLSEKILFEEVGPGQVVTVDVDNWDGEGPGEDAVFTFTGTRKPPAEPDLAKAGAHSAGGPEPAAR
;
2 1 2
;-LDRLTNKFQLALADAQSLALGHDNQFIEPLHLMSALLNQEGGSVSPLLTSAGINAGQLRTDINQALNRLPQVEGT
GGDVQPSQDLVRVLNLCDKLAQKRGDNFISSELFVLAAL-ESRGTLADILKAAGATTANITQAIEQMRGGESVNDQG
AEDQRQ-----------ALKKYTIDLTERAEQGKLDPVIGRDEEIRRTIQVLQRRTKNNPVLIGEPGVGKTAIVEG
LAQRIINGEVPEGLKGRRVLALDMGALVAGAKYRGEFEERLKGVLNDLAKQEGNVILFIDALHTMVGAGKADGAMDA
GNMLKPALARGELHCVGATTLDEYRQYIEKDAALERRFQKVFVAEPSVEDTIAILRGLKERYELHHHVQITDPAIVA
AATLSHRYIADRQLPDKAIDLIDEAASSIRMQIDSKPEELDRLDRRIIQLKLEQQALMKESDEASKKRLDMLNEELS
DKER----QYSELEEEWKAEKASLSGTQTIKAELEQAKIAIEQARRVGDLARMSELQYGKIPELEKQLEAATQLEGK
TMRLLRNKVTDAEIAEVLARWTGIPVSRMMESEREKLLRMEQELHHRVIGQNEAVDAVSNAIRRSRAGLADPNRPIG
SFLFLGPTGVGKTELCKALANFMFDSDEAMVRIDMSEFMEKHSVSRLVGAPPGYVGYEEGGYLTEAVRRRPYSVILL
DAVEKAHPDVFNILLQVLDDGRLTDGQGRTVDFRNTVVIMTSNLGVRETERKSIGLIHQ----DNSTDAM-----EE
IKKIFRPEFINRIDEVVVFHPLGEQHIASIAQIQLKRLYKRLEERGYEIHISDEALKLLSENGYDPVYGARPLKRAI
QQQIENPLAQQILSGELVPGKVIRLEVN-------EDRIVAVQ------------------------
;
```

Local and global model quality metrics: Adapted from ModelArchive entry ma-tbvar3d-18

```
_ma_model_list.ordinal_id                 1
_ma_model_list.model_id                   1
_ma_model_list.model_name                 'Model for P9WPC9'
_ma_model_list.data_id                    4
_ma_model_list.model_type                 'Homology model'

loop_
_ma_qa_metric.id
_ma_qa_metric.name
_ma_qa_metric.description
_ma_qa_metric.type
_ma_qa_metric.mode
1 QMEANDisCo 'Predicted accuracy according to all-atom lDDT in [0,1]' 'pLDDT
all-atom in [0,1]' local
2 QMEANDisCo 'Predicted accuracy according to all-atom lDDT in [0,1]' 'pLDDT
all-atom in [0,1]' global

_ma_qa_metric_global.ordinal_id     1
_ma_qa_metric_global.model_id       1
_ma_qa_metric_global.metric_id      2
_ma_qa_metric_global.metric_value   0.555

loop_
_ma_qa_metric_local.ordinal_id
_ma_qa_metric_local.model_id
_ma_qa_metric_local.label_asym_id
_ma_qa_metric_local.label_seq_id
_ma_qa_metric_local.label_comp_id
_ma_qa_metric_local.metric_id
_ma_qa_metric_local.metric_value
1 1 A 171 LEU 1 0.520
2 1 A 172 ASP 1 0.550
3 1 A 173 GLN 1 0.490
4 1 A 174 PHE 1 0.510
5 1 A 175 GLY 1 0.660
6 1 A 176 ARG 1 0.590
7 1 A 177 ASN 1 0.720
8 1 A 178 LEU 1 0.700
# … abbreviated …
```

References

[1] Westbrook JD, Young JY, Shao C, Feng Z, Guranovic V, Lawson C, et al. (2022). PDBx/mmCIF Ecosystem: Foundational semantic tools for structural biology. J Mol Biol. 434, 167599.

[2] Westbrook J, Henrick K, Ulrich EL, Berman HM. (2005). 3.6.2 The Protein Data Bank exchange data dictionary. In International Tables for Crystallography, (Hall SR, McMahon B, eds), p. 195-198, Springer, Dordrecht, The Netherlands.

[3] Bienert S, Waterhouse A, de Beer TA, Tauriello G, Studer G, Bordoli L, et al. (2017). The SWISS-MODEL Repository-new features and functionality. Nucleic Acids Res. 45, D313-D319.

[4] Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science. 373, 871-876.

[5] Humphreys IR, Pei J, Baek M, Krishnakumar A, Anishchenko I, Ovchinnikov S, et al. (2021). Computed structures of core eukaryotic protein complexes. Science. 374, eabm4805.

[6] Pieper U, Webb BM, Dong GQ, Schneidman-Duhovny D, Fan H, Kim SJ, et al. (2014). ModBase, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res. 42, D336-346.

[7] Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. (2015). The I-TASSER Suite: protein structure and function prediction. Nat Methods. 12, 7-8.

[8] Zheng W, Zhang C, Li Y, Pearce R, Bell EW, Zhang Y. (2021). Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. Cell Rep Methods. 1, 100014.

[9] Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, 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.

[10] Sehnal D, Bittrich S, Velankar S, Koca J, Svobodova R, Burley SK, et al. (2020). BinaryCIF and CIFTools-Lightweight, efficient and extensible macromolecular data management. PLoS Comput Biol. 16, e1008247.

[11] UniProt Consortium. (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 49, D480-D489.

[12] Biasini M, Schmidt T, Bienert S, Mariani V, Studer G, Haas J, et al. (2013). OpenStructure: an integrated software framework for computational structural biology. Acta Crystallographica Series D. 69, 701-709.

[13] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature. 596, 583-589.

[14] Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 46, W296-W303.

[15] Gao M, Coletti M, Davisdon RB, Prout R, Abraham S, Hernandez B, et al. (2022). Proteome-scale Deployment of Protein Structure Prediction Workflows on the Summit Supercomputer. Secondary Proteome-scale Deployment of Protein Structure Prediction Workflows on the Summit Supercomputer, 206-215, doi: 10.1109/IPDPSW55747.2022.00045.

[16] Varadi M, Nair S, Sillitoe I, Tauriello G, Anyango S, Bienert S, et al. (2022). 3D-Beacons: decreasing the gap between protein sequences and structures through a federated network of protein structure data resources. GigaScience. 11, giac118.

[17] Sali A, Blundell TL. (1993). Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 234, 779-815.

[18] Mortuza SM, Zheng W, Zhang C, Li Y, Pearce R, Zhang Y. (2021). Improving fragment-based ab initio protein structure assembly using low-accuracy contact-map predictions. Nature communications. 12, 5011.
[19] Zheng W, Li Y, Zhang C, Zhou X, Pearce R, Bell EW, et al. (2021). Protein structure prediction using deep learning distance and hydrogen-bonding restraints in CASP14. Proteins. 89, 1734-1751.
[20] Zheng W, Wuyun Q, Zhou X, Li Y, Freddolino PL, Zhang Y. (2022). LOMETS3: integrating deep learning and profile alignment for advanced protein template recognition and function annotation. Nucleic Acids Res. 50, W454-W464.
[21] Westbrook JD, Fitzgerald PMD. (2009). Chapter 10 The PDB format, mmCIF formats, and other data formats. In Structural Bioinformatics, Second Edition, (Bourne PE, Gu J, eds), p. 271-291, John Wiley & Sons, Inc., Hoboken, NJ.
[22] Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Zidek A, et al. (2021). Highly accurate protein structure prediction for the human proteome. Nature. 596, 590-596.
[23] Kim DE, Chivian D, Baker D. (2004). Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res. 32, W526-531.
[24] Sehnal D, Bittrich S, Deshpande M, Svobodova R, Berka K, Bazgier V, et al. (2021). Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. Nucleic Acids Res. 49, W431–W437.
[25] Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, et al. (2021). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci. 30, 70-82.