タンパク質立体構造データと紐づけたコーパス作成の試み

佐久間 航也[1,2] 丹羽 智美[3]
[1]名古屋大学 細胞生理学研究センター [2]名古屋大学 高等研究院 [3]大阪大学 蛋白質研究所
echo ksakuma@was@at@i.nagoya-u.ac.jp | sed s/@was@at@/@cesp/g
echo s.niwa@protein.osaka-u.ac.jp | sed s/a@p/@p/g

概要

タンパク質の立体構造データは原子の位置座標を用いて分子の形を表現したものであり、それと同時に、構造生物学者による解釈の対象でもある。AlphaFold などの高精度な立体構造予測手法は、立体構造を計算機で取り扱うことの可能性と価値を証明したが、いまだに立体構造データの解釈、つまり「立体構造が意味するところ」をバイオインフォマティクス的に扱うことは難しい。一方、近年の自然言語処理技術の発展により、言語データを媒体とすることで「意味を計算する」ことが可能となりつつあるようにみえる。したがって、物質科学的な計算手法と自然言語処理を組み合わせることで、分子の機能予測や設計など、物性と意味が絡みあうタイプの科学の発展に寄与することができると期待される。だが、そのような物質科学的データと「それに対する記述」を詳細にペア化したデータセット――物性と意味とをリンクし得る資源――は存在しない。無いなら作れば良い。本稿では、われわれが構造生物学ドメインにおいて進めている立体構造に紐づいたコーパスの要件定義と試作について報告する。

1 本研究の背景

§1 では本研究の背景として必要になる構造生物学の一般的な背景、特にタンパク質の役割や立体構造について簡単にまとめる[1]。

1.1 タンパク質の機能と立体構造

タンパク質は生体分子の一種であり、生命現象における「素子」のような役割を担う。もしあなたが人間で、視覚的に本論文を読んでいるのなら、この瞬間にも網膜にあるロドプシンという光受容タンパク質がフォトダイオードのごとく可視光に応答し、その結果生じた複雑な神経活動を通じてこの文章が認識されている。

このように、生物学的現象の機構を説明しようとすると、おおむね何らかのタンパク質の働きに帰着される。一方、タンパク質の典型的な大きさは差し渡しで数 nm から数十 nm と、通常観測できる生命現象の空間スケールに比べれば非常に小さく、微視的である。したがって、タンパク質が機能する仕組みを原子レベルで理解することで「生命現象を微視的に理解する」こと、つまり「生物システムの挙動を素子レベルで理解する」ことにつながる。

化学的に言えば、タンパク質は多数のアミノ酸残基が重合してできた高分子である。そこで用いられるアミノ酸タイプは基本的に 20 種類で、アラニンなら A、システインなら C、アスパラギン酸なら D … などとアルファベット 1 文字で略記できる。たとえば、ヒトのロドプシンは 348 個のアミノ酸残基が 1 列に連なったもので、その冒頭と終端の各 5 残基は「MNGTE…QVAPA」である。このようなアミノ酸の並びのことを「アミノ酸配列」と呼び、どのタイプのアミノ酸残基がどの順番で並ぶかは、そのタンパク質をコードする遺伝子の DNA 塩基配列により厳密に指定されている。

タンパク質が化合物である以上、化学構造を持つ。使われる 20 種のアミノ酸残基の化学構造は決まっており、それらは常にペプチド結合（アミド結合）を介して 1 本の鎖のように連なるため（図 1A）、アミノ酸配列を指定すれば当該タンパク質の化学構造は一意に指定される。これを「1 次構造」と呼ぶ。

さらにタンパク質は自発的に「特定の立体構造へと折れ畳まる」という一般の高分子にはない性質を持つことが知られている 。この現象をフォールディング（folding）と呼ぶ。つまりタンパク質は、だらりと 1 次元的に伸びた鎖ではなく、勝手に「特定の形」へと収まる特殊な鎖である。この「特定の形」を 3 次構造、あるいは狭い意味での「立体構造」とよぶ（図 1B）。そのタンパク質がどのような立体構

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

造をとるかは、多くの場合、１次構造（アミノ酸配列）によって支配されている。

このような立体構造データを保存する場合、当該分子を構成する原子の原子種と、それら原子の３次元空間における位置座標を集めたもの、つまり「どの原子がどの座標にあるか」を列挙したもので与えるのが一般的である。これらの立体構造データはProtein Data Bank (PDB)に mmCIF 形式（図 1C）ないし PDB 形式のファイルとして統一的に登録され、ユニークな ID（PDB-ID）を付与された上で公開される[2, 3]。公開された構造データは誰でもアクセスでき、ウェブサイトから手元へダウンロードして適当な分子構造描画ソフト（例えば[4]）で眺めることができる。日本からであれば日本蛋白質構造データバンク（PDBj）を通じたアクセスが最も高速である[i]。

2 本研究の方向性

立体構造とテキストをペアにしたデータはこれまで確立されておらず、満たすべき要件も作成手順も未知である。§2 では、このようなデータセットの位置付け、要件、作成の方針を検討する。

2.1 構造生物学という学問分野

構造生物学とは、タンパク質など生体分子の立体構造を実験で明らかにし、それが関与する生命現象を分子レベルで説明する学問領域である。分子の立体構造を明らかにするのは、「分子の『形』が、その機能を説明する」という信念があるためである。この信念を反映し、構造生物学には「構造決定」と「構造の解釈」の２つのフェーズがある。

構造決定に用いられる実験手法としては結晶学、NMR、電子顕微鏡による単粒子解析などがあるが、いずれにしても最終的には分子を構成する原子の位置座標データが得られる。ここまでが構造生物学の「構造」の部分に相当する。

次に、構造生物学者は座標データを生物学的に解釈することを試みる。きわめて単純化してしまえば、構造生物学者は実験で得られた原子座標データを立体物として視覚的に眺める。そして、構造に対して、現在注目している生物学的な謎に関連した所感を述べ、それを論文として報告する。この部分が構造生物学の「生物学」の部分に相当する。

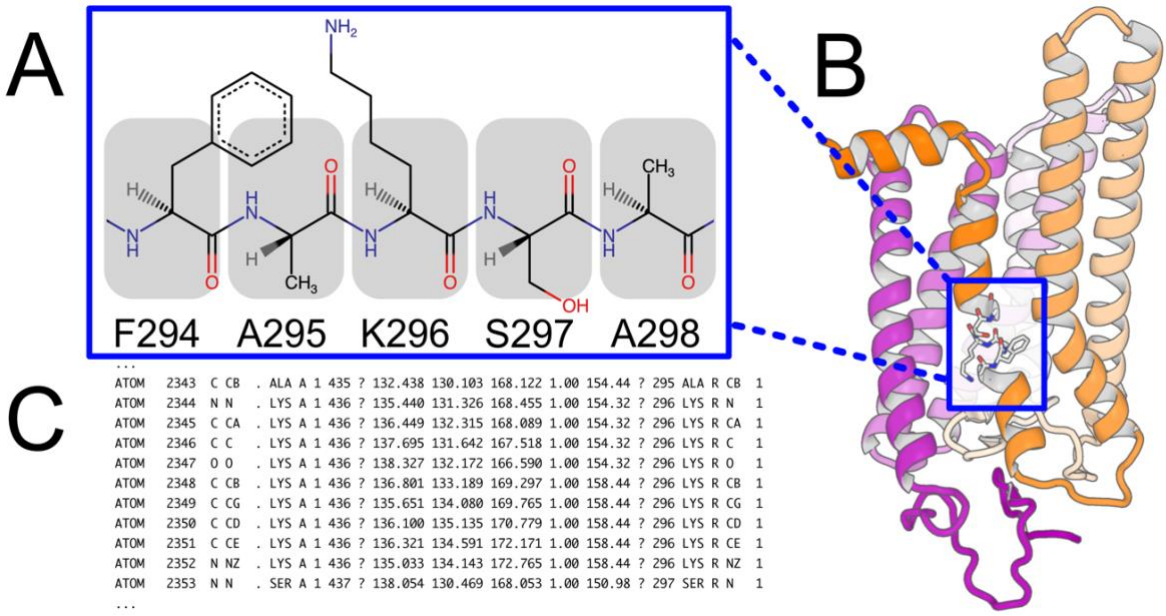

図 1：タンパク質の構造と立体構造を格納するデータ形式

(A) ポリペプチドの化学構造を示す構造式の例：ヒトロドプシンの残基番号 294~298 に相当する一次構造を示した。(B) 3D 物体として描画された立体構造：ヒトロドプシンの立体構造（PDB-ID: 6CMO chain R）をリボンモデルで、残基番号 294～298 のみは(A)の化学構造との対応を示すために stick 表現で描画した。アミノ末端からカルボキシ末端に向けて紫ーオレンジで彩色。 (C)立体構造のデータの例：PDB-ID: 6CMO chain R の mmCIF 形式のデータから K296 に属する全原子と前後の 1 原子の領域を抜き出したもの。4 行目が原子タイプ、6 行目が残基タイプ、11~13 行目が原子座標にあたる。

[i] https://pdbj.org

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

このように構造生物学研究を通じて、「分子の形に対する解釈」が立体構造に紐づいた自然言語の形で与えられる。したがって、PDB に登録された座標データと、その構造を報告している構造生物学論文データをペアとみなすことで、立体構造とその意味を結びつけるための基礎データになると期待される。このような基礎データセットを構築することが本研究の目的である。

2.2 画像・キャプションのアナロジー

座標データと論文での記述の関係性は、画像とキャプションの関係になぞらえると理解しやすい。ただし一般に立体構造データは画像ではないうえ、論文とキャプションには以下のような相違点もある：

(1) テキストが長大である。立体構造に付随する論文は短くても数千ワードからなる。
(2) テキストの焦点が多岐にわたる。論文は必ずしも立体構造データについてのみ記述するわけではなく、通常のキャプションに比べ内容が「散漫」である。
(3) データ数が少ない。実験的に決定された立体構造は 2025 年 1 月 1 日の時点で 229,564 エントリが PDB に登録されており、構造・論文ペア数はこれよりも少ない。
(4) 1 対 1 対応ではない。複数の立体構造を同時に決定して 1 報の論文で言及することがある。
(5) 論文には図表などの画像データも含まれ、純粋なテキストではない。

(1)から(3)の点をふまえると、「立体構造 1 つ」と「論文 1 報」を「1 つのデータペア」と考えるのではなく、立体構造と論文の両方に対して適切なセグメンテーションを行った上でペア化する必要があると我々は考えた。つまり、「特定の部分構造に関する記述」と「その部分構造を指定するマスク情報」を網羅的に抽出してペア化すれば、テキストの焦点が明確化され、さらにその記述が構造上どこに紐づくかも明示でき、データペア数も実質的に増加するため、1 構造-1 論文ペアに比べれば高品質なデータセットになると期待される。 (4)と(5)については、セグメンテーションを前提とした上で、データ形式の工夫で解決できると考えられる。

コストの高いセグメンテーションを正当化する理由として、構造生物学の論文は、立体構造の全体（全体構造）に言及するだけでなく、そこに含まれる様々な要素（部分構造）に対して詳細な説明を与える傾向がある点が挙げられる。たとえば、§4 で取り扱う文献[5]にも見られるように「The MTBD interacted with the hydrophobic core of LC1 mainly via two regions, the H5 helix and the flap region.」といった特定の部分構造についての記述が頻出する。この例では MTBD、the hydrophobic core of LC1、the H5 helix や the flap region が特定の部分構造を指している。したがって、このような局所化された立体構造要素とテキスト要素を紐づけることで空間的・意味的に分解能の高いデータペアが作れると期待できる。

当然、このような記述はテキスト全体の中での文脈と、全体構造の中での「文脈」を考慮することにより意味を持つものだが、文脈中の意味を捉えるためにも局所的な要素同士の対応関係を明らかにする必要がある。また、このような「その立体構造のどの部分に注目するか」ということ自体にも構造生物学者の知見が入るため、これもセグメンテーションによって回収できることを期待する。

以上から、本研究では「実験で決定された構造データ」と「論文テキスト」の組を元データとし、セグメント単位でペアデータを構築することを第一の目標とする。セグメンテーションにはコストがかかるため[ii]、これを迂回できる他のデータ構築方策の可能性も検討するべきだが、これらについては本稿では割愛し[iii]、実験的な構造生物学の知見を論文から抽出・蓄積する方策を見出すことに注力する。

3 対象データの基礎的な統計

本研究の基盤となるデータの状況を把握するため、PDB に登録されている立体構造データと文献について基礎的な調査を行なった。2025 年 1 月 1 日現在、PDB から参照可能なユニークな DOI を持つ文献数は 82,723 であることがわかった。PDB エントリと文献の関係に関するそのほかの調査結果は付録に示した。

[ii] たとえば CLIP[6]では大規模なペアデータを用いることで、あらわにセグメンテーションを行わずとも、テキスト・画像に含まれる要素間の関係を学習できている。これと似たことが実現できればセグメンテーションは不要である。

[iii] UniProt[7]のアノテーションと AlphaFold2[8]による予測立体構造データ[9]から言語・構造ペアを作ることも可能だが、自明な内容であるため割愛する。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

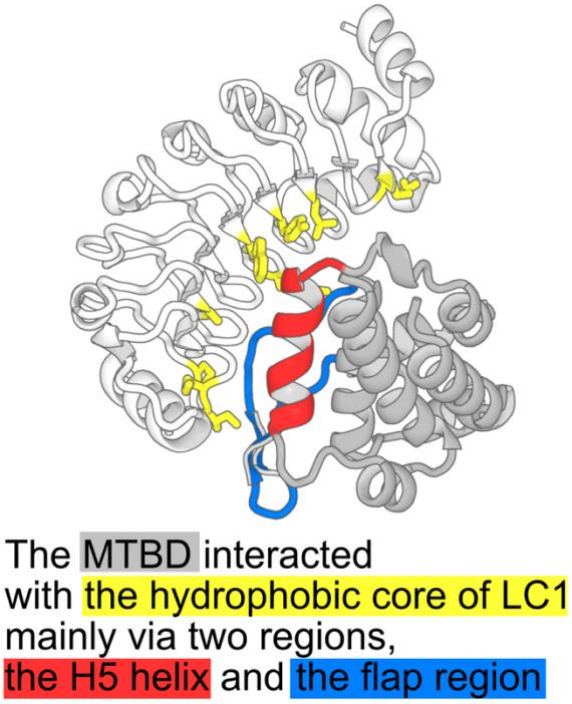

図 2：PDB-ID 6L4P への解釈のマッピング
領域指定子に従い MTBD、the hydrophobic core of LC1、the H5 helix、the flap region を灰色、黄色、赤、青で示した。LC1 全体は白色で示した。

4 試験的アノテーション

データ作成においてどのような状況が発生するか把握するため、実際の文献と PDB エントリを対象に、構造に関する記述と立体構造データに対する領域指定子[iv]のペアの作成を行なった。対象は文献[5]の Results のうち、サブセクション“Overall structure of the LC1-MTBD complex”に絞った。

4.1 基本方針

(1) ある PDB エントリに着目し、その primary citation 文献[v]を記述データのソースとする。
(2) 注目する立体構造についての記述を当該論文からセンテンス単位で改変なく抽出する。
(3) その抽出したセンテンスに含まれる構造要素を、座標データに対する領域指定子として表現する。具体的には鎖番号、残基番号、原子番号の組み合わせで領域指定子を表現する。
(4) (2)にふくまれる構造要素を示す単語などの表現と(3)の領域指定子をペアとして保存する。

4.2 具体例

(1) PDB-ID 6L4P[10]と文献[5]をペアと捉える。
(2) 論文から次のような記述を抽出する。
“The MTBD interacted with the hydrophobic core of LC1 mainly via two regions, the H5 helix and the flap region (Fig. 1B).”
(3) この中に含まれる構造要素と領域指定子を決定する。
a. MTBD = “6L4P and chain B and polymer”
b. The hydrophobic core of LC1 = “6L4P and chain A and resi 33+34+36+54+99+102+121+146+182”
c. the H5 helix = “6L4P and chain B and resi 1734-1746”
d. the flap region = “6L4P and chain B and resi 1689-1708”
(4) これらをまとめる。
pair = {"description": "The MTBD interacted with the hydrophobic core of LC1 mainly via two regions, the H5 helix and the flap region (Fig. 1B)", "element": ["MTBD", "the H5 helix",…], "selection": ["6L4P and chain B and polymer", "6L4P and chain B and resi 1734-1746",…]}

4.3 結果と考察

対象にしたサブセクションの 14 文のうち 13 文が立体構造に言及しており、論文全体ではなく文単位に分割して部分構造と紐づける意味がありそうなことが確認できた。

一方で難しい部分も多数あった。まず、文中で領域指定子が定義されないことがあり、例えば「H5 helix」を指定する残基番号は Supplement の Figure S2 で図示されていた。この場合、テキストの解析のみから領域指定子を得ることができないため、現時点では専門家によるアノテーションの必要性がある。また、テキストデータ部分の言語資源としての価値を高めるには、複数の文に渡る議論を追うための照応解析も必要になると予想される。

構築するデータセットの質を担保していくには、言語資源構築研究の知見および構造生物学的な慣習の両面を考慮してアノテーション内容と規約、データ格納形式を策定していくことが求められる。□

[iv] 画像データに対するマスクに相当する。今回は PyMOL[4]の selection algebra 記法を用いる。

[v] ある PDB エントリの立体構造を初めて報告した文献を、そのエントリに対する primary citation と呼ぶ。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

引用文献

[1] Carl-Ivar Brändén and John Tooze, *Introduction to Protein Structure*. Garland Science, 1999.

[2] H. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank," *Nat. Struct. Mol. Biol.*, vol. 10, no. 12, pp. 980–980, Dec. 2003, doi: 10.1038/nsb1203-980.

[3] J. D. Westbrook *et al.*, "PDBx/mmCIF Ecosystem: Foundational Semantic Tools for Structural Biology," *J. Mol. Biol.*, vol. 434, no. 11, p. 167599, Jun. 2022, doi: 10.1016/j.jmb.2022.167599.

[4] Schrodinger, "The PyMOL Molecular Graphics System, Version 1.8," Nov. 2015.

[5] A. Toda, Y. Nishikawa, H. Tanaka, T. Yagi, and G. Kurisu, "The complex of outer-arm dynein light chain-1 and the microtubule-binding domain of the γ heavy chain shows how axonemal dynein tunes ciliary beating," *J. Biol. Chem.*, vol. 295, no. 12, pp. 3982–3989, Mar. 2020, doi: 10.1074/jbc.RA119.011541.

[6] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," Feb. 26, 2021, *arXiv*: arXiv:2103.00020. Accessed: Nov. 05, 2024. [Online]. Available: http://arxiv.org/abs/2103.00020

[7] The UniProt Consortium *et al.*, "UniProt: the Universal Protein Knowledgebase in 2023," *Nucleic Acids Res.*, vol. 51, no. D1, pp. D523–D531, Jan. 2023, doi: 10.1093/nar/gkac1052.

[8] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021, doi: 10.1038/s41586-021-03819-2.

[9] M. Varadi *et al.*, "AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences," *Nucleic Acids Res.*, vol. 52, no. D1, pp. D368–D375, Jan. 2024, doi: 10.1093/nar/gkad1011.

[10] Toda, A., Nishikawa, Y., Tanaka, H., Yagi, T., Kurisu, G., "Crystal structure of the complex between the axonemal outer-arm dynein light chain-1 and microtubule binding domain of gamma heavy chain." Feb. 19, 2020. doi: https://doi.org/10.2210/pdb6l4p/pdb.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

概要

PDB から参照可能な primary citation の統計

2025 年 1 月 1 日時点での PDB 全エントリ 229,564 件のうち、DOI を持った primary citation 文献と紐づいているものは 190,718 件（83%）であり、さらに PubMedID を持つものは 187,154 件（81%）であった。逆に、ユニークな文献 DOI は 82,723 であり、平均して 1 文献で 2.3 構造が報告されていた。1 文献で同時に報告された PDB エントリを Related PDB Entry と呼ぶ。Related PDB Entry 件数の統計から、4 構造を同時に報告した論文まで対象にいれれば全文献の 90%をカバーできることがわかる（表 1）。

表 1：同一 Primary Citation を持つ PDB エントリ数の統計

Number of Related PDB Entries (max 860)	文献数	累積相対度数
1	41,770	0.504
2	19,025	0.734
3	9,107	0.845
4	5,032	0.905
5	2,731	0.938
6	1,650	0.958

データ形式の素案

体系的なデータ蓄積を進めるためには、テキストと座標データを統一的に保持するためのデータ形式の策定が必要になる。我々は PDBx/mmCIF 形式[1]にテキストアノテーションを取り込めるデータブロックを追加して拡張することを検討している。まず mmCIF 形式は構造生物学分野で広く使われているため、構造データを無理なく収録できるデータ形式であり各種のパーサー開発や分子構造描画ソフトの対応も進んでいる。また、PDBx/mmCIF 形式は JSON（PDBx/mmJSON[2, 3])形式と相互変換できることも大きな利点である。テキストを含む本研究のデータセットが下流でデータ科学的に利用されることを念頭におけば、このような汎用データ形式と行き来できるのは大きな利点になると考えられる。以下に mmCIF 形式を拡張してテキストアノテーションを含められるようにする方策の素案を示し、暫定的に AnnotCIF と呼称する。記載内容は完全に架空のものであることに注意。

```
# AnnotCIF Example v0.0.1 #
data_0XYZ
loop_
_annotator_list.id
_annotator_list.name
_annotator_list.affiliation
_annotator_list.orcid
_annotator_list.email
_annotator_list.role
1 "John Smith" "Prefectural University of Nagoya" "0000-0001-2345-XXXX" "jsmith.at.example.dagaya.ac.jp" "curator"
2 "Jane Smith" "Nippon National Institute for Structural Bioinformatics" "0000-0002-3456-XXXX" "smithj.at.example.nnisbi.ac.jp" "validator"

loop_
_struct_segment.id
_struct_segment.auth_begin_id
_struct_segment.auth_end_id
_struct_segment.auth_asym_id
_struct_segment.auth_atom_id
_struct_segment.label_begin_id
_struct_segment.label_end_id
_struct_segment.label_asym_id
_struct_segment.label_atom_id
_struct_segment.segment_type
1  23  35  A  ALL 23  35  A  ALL "beta_sheet"
2  40  52  A  ALL 40  52  A  ALL "alpha_helix"
3  15  15  A  ALL 15  15  A  ALL "active_site"
3  60  65  A  ALL 60  65  A  ALL "active_site"
3  102 102 A  ALL 102 102 A  ALL "active_site"
3  30  35  B  ALL 30  35  B  ALL "active_site"
4  70  85  A  ALL 70  85  A  ALL "binding_pocket"
5  120 125 A  ALL 120 125 A  ALL "metal_binding"
5  150 150 A  ALL 150 150 A  ALL "metal_binding"
5  205 205 A  ZN  1   1   C  ZN  "ligand_metal"

loop_
_struct_segment_annotation.id
_struct_segment_annotation.segment_id
_struct_segment_annotation.annotator_id
_struct_segment_annotation.source_text
_struct_segment_annotation.source_doi
_struct_segment_annotation.confidence_score
_struct_segment_annotation.timestamp
1 1 1 "Residues 23-35 of chain A form a beta sheet that participates in substrate recognition" "00.0000/j.structure.2025.01.001" 0.9 "2024-03-15T14:30:00"
2 2 1 "The alpha helix undergoes significant rearrangement when the ligand is bound" "00.0000/j.structure.2025.01.001" 0.8 "2024-03-16T09:15:00"
3 3 1 "The catalytic pocket is formed by residues A15, A60-65, A102 and B30-35" "00.0000/j.structure.2023.01.001" 0.95 "2025-03-15T16:45:00"
4 4 1 "Residues 70-85 forming the binding pocket are highly conserved in the protein family" "00.0000/j.structure.2025.01.001" 0.85 "2024-03-17T10:20:00"
5 5 2 "Residues 120-125 and 150 coordinate a zinc ion essential for structural stability" "00.0000/j.structure.2025.01.001" 0.9 "2024-03-18T11:30:00"

loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_alt_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_entity_id
_atom_site.label_seq_id
_atom_site.pdbx_PDB_ins_code
_atom_site.Cartn_x
----------------------------------------------------------------------------TRUNCATED----------------------------------------------------------------------------
```


[1] J. D. Westbrook *et al.*, “PDBx/mmCIF Ecosystem: Foundational Semantic Tools for Structural Biology,” *Journal of Molecular Biology*, vol. 434, no. 11, p. 167599, Jun. 2022, doi: 10.1016/j.jmb.2022.167599.

[2] G.-J. Bekker, H. Nakamura, and A. R. Kinjo, “Molmil: a molecular viewer for the PDB and beyond,” *J Cheminform*, vol. 8, no. 1, p. 42, Dec. 2016, doi: 10.1186/s13321-016-0155-1.

[3] A. R. Kinjo *et al.*, “New tools and functions in data-out activities at Protein Data Bank Japan (PDBj),” *Protein Science*, vol. 27, no. 1, pp. 95–102, Jan. 2018, doi: 10.1002/pro.3273.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).