La molécule du mois : fêter les 55 ans de la banque de données sur les protéines (PDB)

Publié le 31.10.26 | Par Janet Iwasa

Une banque de données ayant permis l'évolution rapide de la prédiction et de la conception de structures de protéines.

Depuis le cinquantenaire de la banque de données sur les protéines (PDB, *Protein Data Bank* en anglais) en 2021, le principal développement en biologie structurale a été l'essor des structures déterminées par ordinateur, que ce soit par AlphaFold2, RoseTTAFold ou d'autres programmes de prédiction de structures de protéines reposant sur des approches d'intelligence artificielle et d'apprentissage automatique. Cet article retrace l'évolution de la prédiction des protéines à travers l'exploration de trois structures qui reposent sur le développement de ces nouvelles méthodes de calcul : la prédiction réussie d'une protéine unique, la modélisation d'un complexe moléculaire flexible et la création d'un bactériophage issu d'un génome synthétique.

1. La banque de données sur les protéines comme socle

Les méthodes actuelles de prédiction de protéines ont été rendues possibles par la richesse des données structurales déjà contenues dans la banque de données sur les protéines. AlphaFold et les autres algorithmes peuvent être considérés comme des systèmes de reconnaissance de formes, qui ont appris à reconnaître des régularités (comme la façon dont une séquence d'acides aminés détermine un repliement ou dont différentes chaînes peuvent s'assembler) grâce aux plus de 150 000 structures déterminées expérimentalement que contenait l'archive de la PDB au moment de leur entraînement.

2. Le défi des CASP

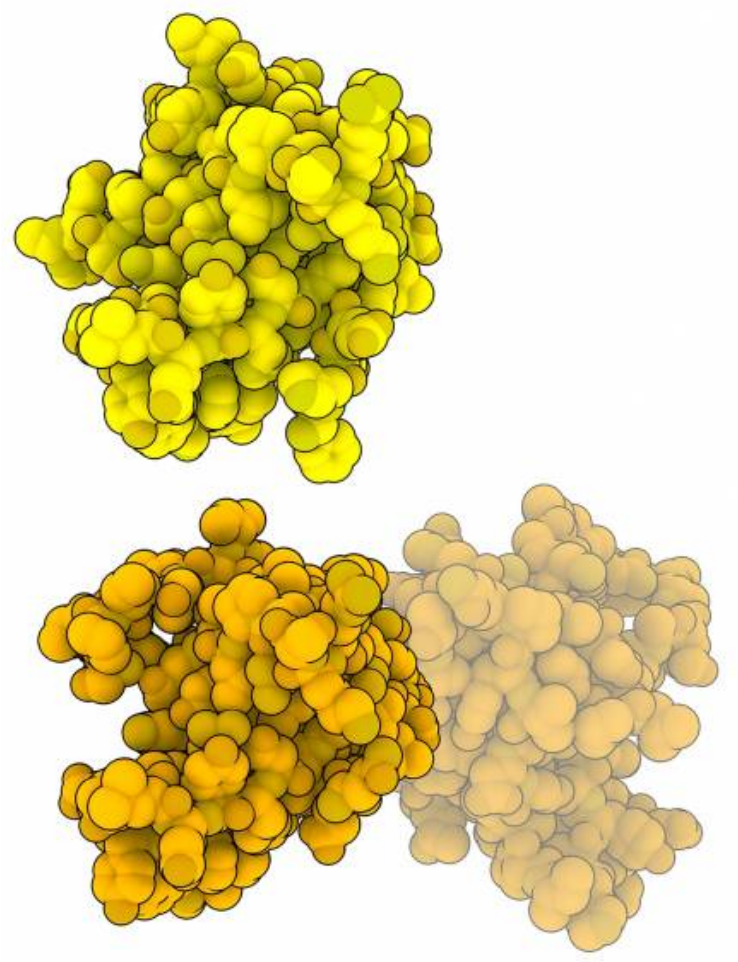


Figure 1 - Comparaison entre la structure de la protéine ORF8 du SARS-CoV-2 prédite par AlphaFold2 et celle déterminée expérimentalement
La structure d'ORF8 du SARS-CoV-2 prédite par l'équipe d'AlphaFold2 et soumise au concours CASP 14 est représentée en jaune. La structure obtenue expérimentalement (entrée PDB 7jtl), qui a été révélée par cristallographie aux rayons X sous forme de dimère, est représentée en orange.

Auteur(s)/Autrice(s) : Janet Iwasa et RCSB PDB – Licence : CC-BY – Source : RCSB PDB

Organisée pour la première fois en 1994, la CASP (*Critical Assessment of protein Structure Prediction* en anglais, ou évaluation critique de la prédiction de structures protéiques) est un concours biennal qui met la communauté scientifique au défi de prédire la structure tridimensionnelle de protéines à partir de leur séquence. Plus la structure prédite [*par un modèle*] est proche de la structure obtenue expérimentalement, plus le score attribué [*à ce modèle*] est élevé. Pendant des décennies, la précision moyenne des prédictions réalisées dans le cadre des CASP est restée autour de 20 %. Avec l'utilisation croissante de l'apprentissage automatique et des réseaux de neurones profonds, les scores ont nettement augmenté lors des CASP 12 et 13 (organisées en 2016 et 2018), tout en laissant une marge de progression importante. Lors de la CASP 14, organisée en 2020, un net favori est cependant apparu : AlphaFold2, de DeepMind, qui a produit de façon constante des structures dont la précision était comparable à celle de techniques expérimentales de plus faible résolution **[1] [2] [3]**.

L'une des protéines cibles de la CASP 14 était la protéine ORF8 du SARS-CoV-2. ORF8 est une petite protéine accessoire du virus, à évolution rapide, et qui atténue la réponse immunitaire de l'hôte. La structure d'ORF8 était difficile à prédire car sa séquence ressemble très peu à celle d'autres protéines dont la structure a été déterminée expérimentalement. Comme cela est représenté sur la figure 1, AlphaFold2 a produit une prédiction précise de la structure expérimentale d'ORF8 (entrée PDB 7jtl) **[4]**, déterminée par la suite.

3. Résoudre la structure d'un complexe flexible

Quelques années seulement après le succès d'AlphaFold2, les structures de protéines prédites par ordinateur sont désormais couramment utilisées pour aider à déterminer les structures d'assemblages de plusieurs sous-unités. Un exemple de cette

évolution est le complexe augmine, ou HAUS. L'augmine est un complexe de huit sous-unités qui contribue à la construction du fuseau mitotique en se fixant sur les côtés des microtubules et en induisant la nucléation de nouveaux microtubules ramifiés grâce au recrutement de la γ-tubuline. Sa forme allongée et sa flexibilité la rendaient toutefois difficile à résoudre par les seules méthodes de cristallographie ou de cryo-microscopie électronique de particules uniques.

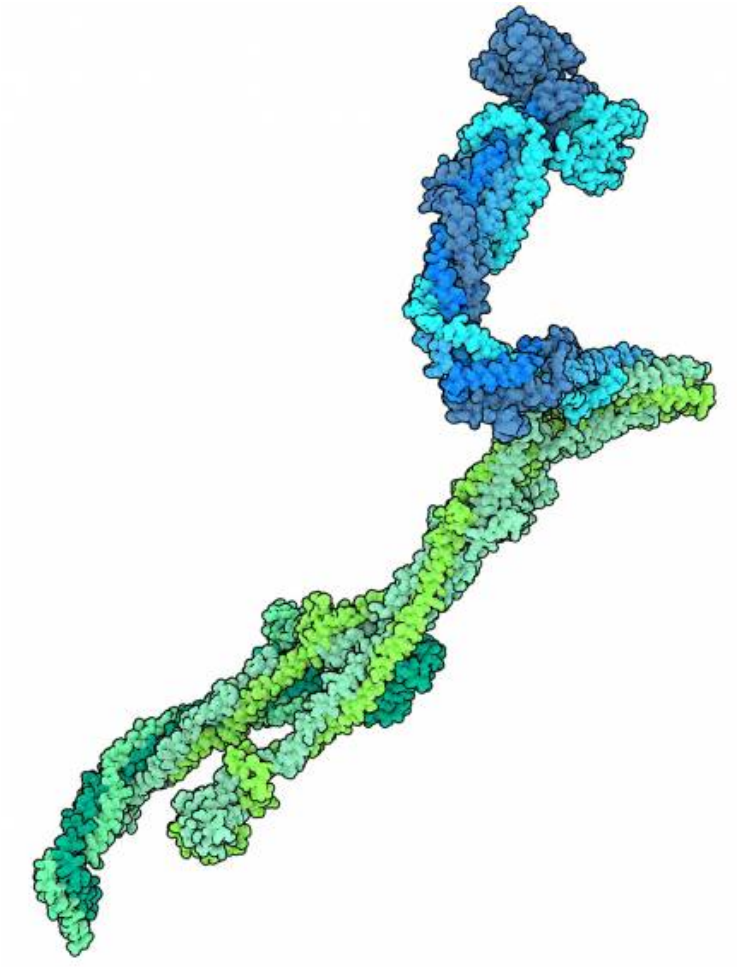


Figure 2 - La structure de l'augmine, résolue par une approche intégrative
La structure de l'augmine (entrée PDB 7sqk) a été résolue par une approche intégrative comprenant des méthodes de prédiction de structure.

Auteur(s)/Autrice(s) : Janet Iwasa et RCSB PDB – Licence : CC-BY – Source : RCSB PDB

Récemment, des chercheurs ont pu résoudre la structure du complexe augmine (représentée sur la figure 2, entrée PDB 7sqk) grâce à une approche intégrative, ou hybride. Des sous-unités individuelles du complexe augmine et de petits groupes de protéines ont été modélisés avec AlphaFold2, puis ajustés à des structures déterminées expérimentalement par cryo-microscopie électronique, avec l'appui de données supplémentaires de pontage chimique couplé à la spectrométrie de masse[1]. La structure obtenue révèle un complexe très flexible de 45 nanomètres de long, avec une tête en forme de V, que les scientifiques pensent impliquée dans la liaison aux microtubules, et une queue bifurquée probablement impliquée dans le recrutement de la γ-tubuline [5].

4. Générer un bactériophage

Le dernier exemple présenté repose sur une utilisation très différente du calcul et de l'IA. Au lieu de prédire une structure à partir d'une séquence connue, les chercheurs ont généré par calcul de nouveaux génomes de bactériophages à l'aide de modèles de langage de génomes [*entraînés sur plus de 2 millions de génomes de bactériophages. Les modèles de langage ont ensuite été interrogés de manière à générer des génomes ayant pour modèle*] le bactériophage ΦX174, un petit virus icosaédrique dont le génome, constitué d'ADN simple brin, comporte onze gènes. Parmi de nombreux génomes candidats générés, l'équipe a

synthétisé et testé près de trois cents génomes conçus par ordinateur et a obtenu seize bactériophages viables, capables d'infecter des bactéries résistantes à ΦX174. L'un des bactériophages générés, Evo-Φ36 (entrée PDB 36cr), est représenté sur la figure 3. Evo-Φ36 est très semblable au modèle ΦX174 (entrée PDB 36cq), mais présente des changements importants dans sa version de la protéine J, une petite protéine interne qui aide à empaqueter le génome et qui, de l'intérieur, soutient la capside. Comme représenté sur la figure 3, Evo-Φ36 a incorporé une version plus courte de la protéine J, [*qui est issue d'un autre bactériophage, appelé G4*], éloigné de ΦX174 d'un point de vue évolutif. À lui seul, ce changement aurait conduit à un bactériophage non viable. Le génome généré portait cependant aussi un ensemble de modifications associées, qui ont rendu ce changement de la protéine J compatible avec sa capside [6].

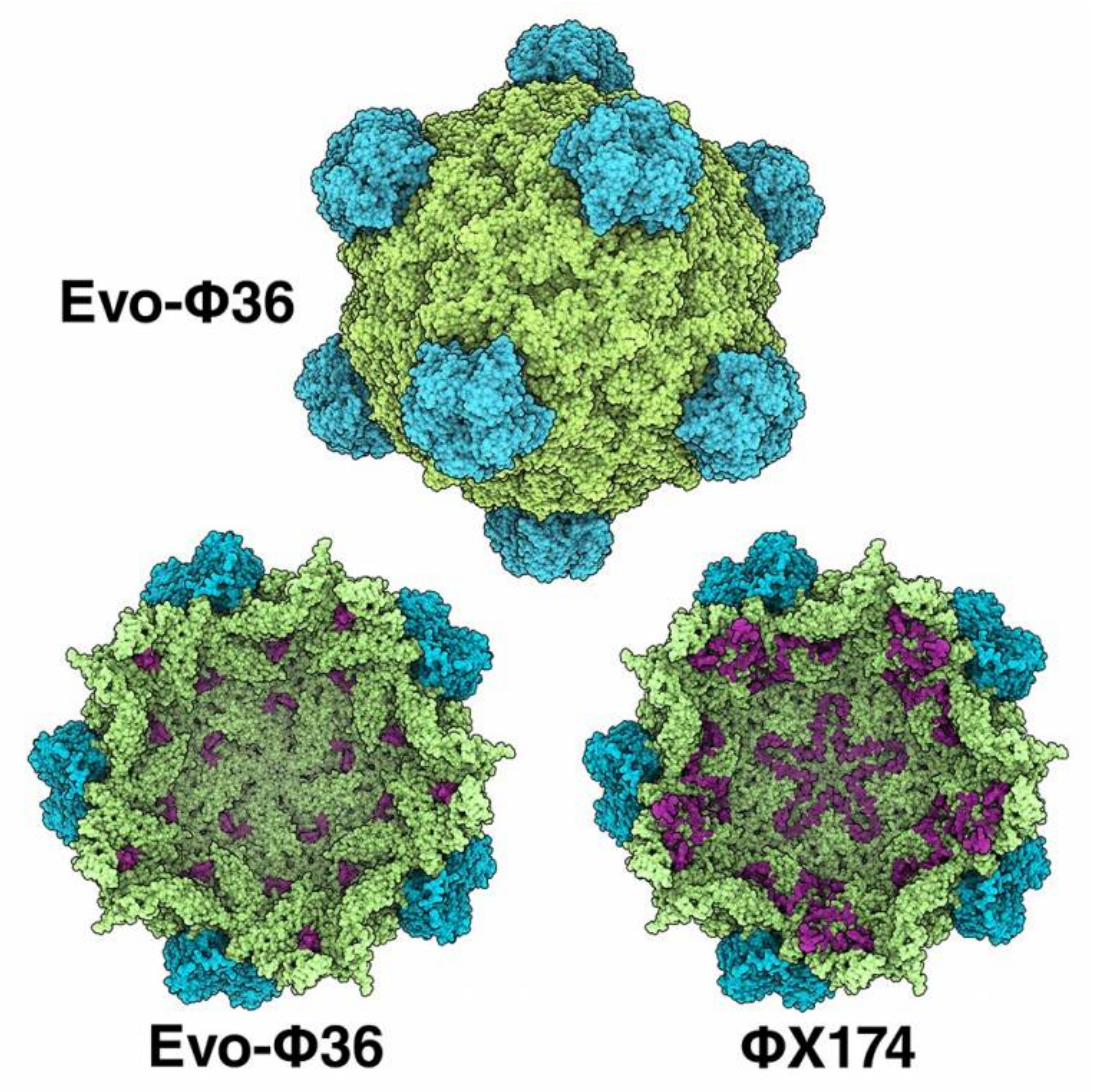


Figure 3 - Le bactériophage synthétique Evo-Φ36
Le bactériophage synthétique Evo-Φ36 est représenté en haut et en bas à gauche (entrée PDB 36cr), tandis que le bactériophage ayant servi de référence, ΦX174 (entrée PDB 36cq), est représenté en bas à droite. Les structures du bas montrent la capside virale en coupe. La protéine J est représentée en violet.

Auteur(s)/Autrice(s) : Janet Iwasa et RCSB PDB – Licence : CC-BY – Source : RCSB PDB

5. Défis restants

Si le succès des méthodes de prédiction de protéines est indéniable, des défis demeurent, comme l'a souligné la CASP17 [en 2026]. Ils concernent notamment la prédiction des complexes protéine-ligand, des acides nucléiques et des complexes protéine-acide nucléique, ainsi que des ensembles conformationnels, qui décrivent la gamme d'états qu'une molécule ou un complexe explore au cours du temps.

6. Explorer la structure

Vous pouvez comparer des structures obtenues expérimentalement avec les prédictions de groupes très bien classés aux CASP 10 (organisé en 2012), CASP 12 (2016) et CASP 14 (2020). Les structures prédites sont représentées en jaune et les structures déterminées expérimentalement en orange. Chacune de ces structures avait été placée dans la catégorie « difficile », car elle ne présentait que peu de similarité structurale détectable avec une protéine déjà résolue dans la banque de données sur les protéines. Les protéines mises en avant sont : pour la CASP 10, RUMGNA_01417 (entrée PDB 3nrl) et une prédiction du groupe de Zhang ; pour la CASP 12, la protéase monomérique pUL26N du virus de la pseudo-rage (entrée PDB 4cx8) et une prédiction du groupe de Baker ; pour la CASP 14, ORF8 du SARS-CoV-2 (entrée PDB 7jtl) et une prédiction du groupe AlphaFold2. Vous pouvez comparer ces structures en vous rendant sur la page de l'article d'origine (en anglais), et en cliquant sur l'onglet JSmol de la section *Exploring the Structure*.

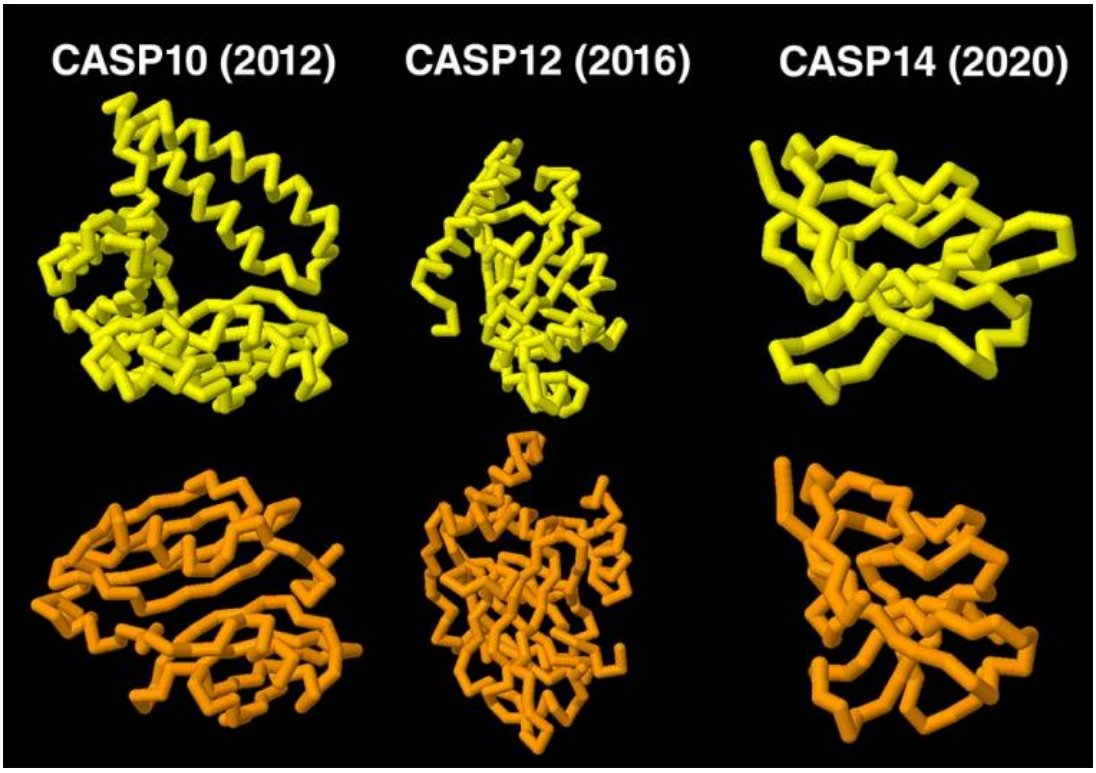


Figure 4 - Comparaison, pour trois protéines, des structures prédites et des structures déterminées expérimentalement

La comparaison des structures prédites au cours de différents concours CASP (en haut, en jaune) avec les structures déterminées expérimentalement (en bas, en orange). À gauche, la protéine RUMGNA_01417 de la bactérie pathogène *Ruminococcus gnavus*, au centre la protéase pUL26N du virus de la pseudo-rage, et à droite la protéine ORF8 du SARS-CoV-2.

Auteur(s)/Autrice(s) : Janet Iwasa et RCSB PDB – Licence : CC-BY – Source : RCSB PDB

7. Pour aller plus loin

1. Vous pouvez en apprendre davantage sur les protéines conçues sur mesure (*designer proteins*), créées à partir de principes biologiques.
2. Vous pouvez lire les articles célébrant le cinquantenaire et les 40 ans de la banque de données sur les protéines.

Ce texte correspond à la traduction par Cédric Bordi de l'article Molecule of the Month : Celebrating 55 Years of the PDB **[7]** écrit par Janet Iwasa en octobre 2026 sur le site PDB-101, le portail éducatif de la banque de données sur les protéines (PDB). Les ajouts entre crochets ainsi que la note de bas de page correspondent à des précisions apportées par le traducteur.

CRÉDITS

AUTEUR(S)/AUTRICE(S)

Janet Iwasa

Janet Iwasa est professeure assistante au département de biochimie de l'université de l'Utah (États-Unis). Elle est spécialiste de la création d'images et d'animations permettant de rendre compte des processus biologiques.

LICENCE DU TEXTE DE L'ARTICLE

REPRODUIT AVEC AUTORISATION

NOTES

1 Le pontage chimique couplé à la spectrométrie de masse (*crosslinking mass spectrometry*) est une technique permettant notamment de déterminer les interactions entre les sous-unités d'un complexe protéique. L'utilisation d'un agent pontant permet de former des liaisons covalentes entre des acides aminés proches dans l'espace, et donc notamment entre des acides aminés appartenant à différentes sous-unités (rappelons que, *in vivo*, les sous-unités d'un complexe protéique n'interagissent entre elles que par des interactions faibles, non covalentes). Les complexes protéiques pontés sont ensuite digérés en peptides, qui sont alors étudiés par spectrométrie de masse. L'analyse bio-informatique des données permet alors de contraindre les distances entre acides aminés pontés et donc notamment de savoir quelles sous-unités sont proches dans l'espace. Le pontage chimique couplé à la spectrométrie de masse permet donc de compléter les informations apportées par d'autres types d'approches : données de cryo-microscopie électronique ou de cristallographie aux rayons X, structures prédites par intelligence artificielle...

BIBLIOGRAPHIE

1. Jumper, J. ., Evans, R. ., Pritzel, A. ., Green, T. ., Figurnov, M. ., Ronneberger, O. ., ... Hassabis, D. . (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*, 583–589. http://doi.org/10.1038/s41586-021-03819-2 (Original work published 2026)

2. Jumper, J. ., Evans, R. ., Pritzel, A. ., Green, T. ., Figurnov, M. ., Ronneberger, O. ., ... Hassabis, D. . (2021). Applying and improving AlphaFold at CASP14. *Proteins, 89*, 1711–1721. http://doi.org/10.1002/prot.26257 (Original work published 2026)

3. Baek, M. ., DiMaio, F. ., Anishchenko, I. ., Dauparas, J. ., Ovchinnikov, S. ., Lee, G. R., ... Baker, D. . (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science, 373*, 871–876. http://doi.org/10.1126/science.abj8754 (Original work published 2026)

4. Flower, T. G., Buffalo, C. Z., Hooy, R. M., Allaire, M. ., Ren, X. ., & Hurley, J. H. (2021). Structure of SARS-CoV-2 ORF8, a rapidly evolving immune evasion protein. *Proceedings of the National Academy of Sciences of the United States of America, 118*, e2021785118. http://doi.org/10.1073/pnas.2021785118 (Original work published 2026)

5. Gabel, C. A., Li, Z. ., DeMarco, A. G., Zhang, Z. ., Yang, J. ., Hall, M. C., ... Chang, L. . (2022). Molecular architecture of the augmin complex. *Nature Communications, 13*, 5449. http://doi.org/10.1038/s41467-022-33227-7 (Original work published 2026)

6. King, S. H., Driscoll, C. L., Li, D. B., Guo, D. ., Merchant, A. T., Brixi, G. ., ... Hie, B. L. (2026). Generative design of bacteriophages with genome language models. *Science, 393*, eaec2657. http://doi.org/10.1126/science.aec2657 (Original work published 2026)

7. Iwasa, J. . (2026). celebrating 55 years of the PDB. *RCSB Protein Data Bank*. http://doi.org/10.2210/rcsb_pdb/mom_2026_10 (Original work published 2026)