DOI: 10.34749/3061-1466.2026.3

CavitOmiX: A Proteome-Wide AI Framework for Structure-Based Off-Target Prediction, Drug Repurposing, and Antiviral Discovery

Hetmann M., Lena Parigger, Tobias Schopper, Markus Fleck, Georg Steinkellner, Christian C. Gruber
Innophore GmbH, 8010, Graz, Austria

Abstract

Polypharmacology — the unintended binding of drug candidates to non-target proteins — remains a leading cause of late-stage clinical trial failure, yet it is largely invisible to conventional sequence-based screening methods. We present CavitOmiX, an AI-driven computational framework that addresses this problem through proteome-wide *cavitome* analysis: systematic mapping of all potential drug-binding sites across entire proteomes as high-dimensional physicochemical point cloud fingerprints. Leveraging GPU-accelerated protein structure prediction (AlphaFold2, ESMFold, OpenFold) and Innophore's Catalophore™ fingerprinting technology, CavitOmiX enables proteome-scale nearest-neighbour cavity searches. The human cavitome, comprising over 800,000 binding site fingerprints derived from 122,907 AI-predicted structural models (published open-source in *Nature Scientific Data*), serves as a reference for early off-target liability prediction. The same infrastructure supports drug repurposing by matching new targets against approved drug binding sites, and antiviral lead optimization via a genetic algorithm. We validate the framework across three use cases: resistance monitoring in SARS-CoV-2, structural proteome analysis of Monkeypox virus, and full-pipeline antiviral design for Chikungunya virus. All functionality is accessible via a public web application at `copilot.cavitomix.bio`.

1 Introduction

Early-stage drug discovery relies heavily on identifying compounds that bind selectively to a desired target protein while avoiding unintended interactions with other proteins in the human body. Detecting such off-target interactions computationally is non-trivial: protein binding site geometry is evolutionarily far more conserved than amino acid sequence, meaning two proteins with as little as 8% sequence identity can share a nearly identical binding pocket (1). Sequence-based screening — still the dominant paradigm for large-scale virtual screening — systematically misses this structural convergence, leaving significant off-target liabilities undetected until costly biochemical or clinical testing.

The same structural reasoning motivates drug repurposing: if an approved drug's target shares binding site geometry with a new disease target, that drug is a direct repurposing candidate — regardless of any sequence similarity between the two proteins.

The widespread availability of AI-based protein structure prediction models (2) has, for the first time, made it computationally feasible to map binding sites across entire proteomes rather than individual proteins. CavitOmiX was built to exploit this opportunity: it computes, stores, and queries proteome-wide binding site fingerprints at a scale and speed that enables practical integration into early drug discovery workflows.

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

2 Methods

Proteome-Wide Structure Prediction. 3D structural models are generated for viral and human proteomes using a hybrid pipeline of AlphaFold2 (2), ESMFold (3), and OpenFold (4), complemented by physics-based homology modelling for low-confidence regions (5; 6). For the human proteome, inference was performed on NVIDIA BioNeMo GPU infrastructure (7). The resulting dataset of 122,907 non-redundant human structural models was released as an open-source benchmark resource in *Nature Scientific Data* (6).

Cavitome Fingerprinting. Binding-competent cavities are identified across all structural models using geometric cavity detection. Each cavity is encoded as a multidimensional point cloud fingerprint using Innophore's Catalophore™ technology, capturing up to 30 physicochemical and geometric properties per binding site — including hydrogen bond potential, electrostatic surface, hydrophobicity, and shape descriptors — independently of the underlying amino acid sequence (8). The resulting human cavitome contains over 800,000 fingerprints in a queryable vector store.

Proteome-Scale Off-Target Search. Given a query binding site (e.g., a viral drug target), a nearest-neighbour search against the human cavitome identifies structurally similar human proteins at 5,000,000 comparisons per second (8). This represents a 1,250,000-fold speedup compared to iterative closest point (ICP)-based cavity alignment, making real-time proteome-wide queries practical for the first time.

Drug Repurposing via Cavitome Matching. A reference database of binding site fingerprints for approved drug targets is queried in parallel. Structural similarity between a new target and an existing approved drug target surfaces repurposing candidates independently of any sequence relationship between the two proteins (8).

3 Results

Monkeypox Virus Structural Proteome. The structure prediction pipeline was applied to the Monkeypox virus proteome, producing 570 high-confidence structural models — the most comprehensive open-source structural dataset for this pathogen to date (5). Cavitome matching against the human reference database identified a subset of Monkeypox targets with low off-target liability in the human proteome, prioritising them as candidate drug targets.

Human Cavitome Open-Source Release. The full human structural proteome (122,907 models) and derived cavitome fingerprints are publicly available (6), providing the community with a sequence-independent off-target reference for computational drug safety profiling and a large-scale benchmark for structure-based machine learning models.

Chikungunya Virus — Full Pipeline Validation. An end-to-end demonstration was performed for Chikungunya virus (CHIKV), a re-emerging arbovirus with no approved antiviral treatment (8). Starting from CHIKV genomic sequences, structural models of all viral proteins were generated, cavitome fingerprints computed, and repurposing candidates identified via cavity-based matching against approved drug targets. A lead compound was subsequently optimized over 50 evolutionary generations, yielding a candidate with enhanced predicted binding to the CHIKV nsP2 protease and reduced predicted affinity to structurally homologous human kinases — demonstrating selectivity improvement without manual medicinal chemistry iteration.

Web Application. All framework functionality is accessible through CavitOmiX Copilot (`copilot.cavitomix.bio`), a web application that accepts a target protein or binding site as input and returns ranked off-target candidates from the human cavitome alongside structural repurposing suggestions. The tool is actively used by research groups worldwide.

4 Discussion

CavitOmiX addresses a structural blind spot in computational drug discovery by replacing sequence-based similarity with geometry- and physicochemistry-based cavity comparison. The key contributions are: (i) the first proteome-scale, sequence-independent off-target reference database for the human proteome; (ii) a 1.25-million-fold acceleration of cavity-based searches enabling real-time

proteome-wide queries; (iii) a validated end-to-end pipeline from genomic sequence to optimized lead compound; and (iv) an open-access web interface lowering the barrier to adoption.

From an AI perspective, the pipeline integrates multiple machine learning modalities: generative structure prediction (AlphaFold2/ESMFold), high-dimensional vector retrieval, and evolutionary optimization. The open-source human structural dataset (6) further provides a large-scale supervised learning resource for future structure-based models.

The framework is not limited to antiviral applications. Any therapeutic indication in which small molecules target protein binding sites — oncology, neurodegenerative diseases, antimicrobial resistance — can benefit from cavitome-based off-target screening and repurposing analysis.

5 Conclusion

CavitOmiX provides a practical, scalable infrastructure for proteome-wide binding site analysis that integrates seamlessly into early drug discovery. By enabling rapid off-target profiling, structure-based repurposing, and evolutionary lead optimization from genomic sequence input, the framework has the potential to reduce late-stage clinical attrition and accelerate response to emerging infectious diseases. The publicly available dataset, web application, and open codebase invite community adoption and extension.

References

[1] Hetmann, M., Langner, C., Durmaz, V., Cespugli, M., Köchl, K., Krassnigg, A., Blaschitz, K., Groiss, S., Loibner, M., Ruau, D., Zatloukal, K., Gruber, K., Steinkellner, G., Gruber, C.C. (2023). Identification and validation of fusidic acid and flufenamic acid as inhibitors of SARS-CoV-2 replication using DrugSolver CavitomiX. *Scientific Reports*, 13(1), 11783.

[2] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596, 583–589.

[3] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379, 1123–1130.

[4] Ahdritz, G., Bouatta, N., Floristean, C., Kadyan, S., Xia, Q., Gerecke, W., O'Donnell, T.J., Berenberg, D., Fisk, I., Zanichelli, N., Zhang, B., Nowaczynski, A., Wang, B., Stepniewska-Dziubinska, M.M., Zhang, S., Ojewole, A., Guney, M.E., Biderman, S., Watkins, A.M., Ra, S., Lorenzo, P.R., Nivon, L., Weitzner, B., Ban, Y.A., Chen, S., Zhang, M., Li, C., Song, S.L., He, Y., Sorger, P.K., Mostaque, E., Zhang, Z., Bonneau, R., AlQuraishi, M. (2024). OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nature Methods*, 21(8), 1514–1524.

[5] Parigger, L., Krassnigg, A., Grabuschnig, S., Gruber, K., Steinkellner, G., Gruber, C.C. (2023). AI-assisted structural consensus-proteome prediction of human monkeypox viruses. *Microbiology Spectrum*, 11(6), e02315-23.

[6] Hetmann, M.†, Parigger, L.†, Sirelkhatim, H., Stern, A., Krassnigg, A., Gruber, K., Steinkellner, G., Ruau, D., Gruber, C.C. (2024). Folding the human proteome using BioNeMo: A fused dataset of structural models for machine learning purposes. *Scientific Data*, 11(1), 591. († equal contribution)

[7] Costa, A.B., Tadimeti, N., Tretina, K., et al. (2024). BioNeMo Framework: A Modular, High-performance Library for AI Model Development in Drug Discovery. *arXiv preprint*, arXiv:2411.10548.

[8] Parigger, L., Krassnigg, A., Hetmann, M., Hofmann, A., Gruber, K., Steinkellner, G., Gruber, C.C. (2024). CavitOmiX Drug Discovery: Engineering Antivirals with Enhanced Spectrum and Reduced Side Effects for Arboviral Diseases. *Viruses*, 16(8), 1186.