УДК 004.8:57

Foundational Models in Biology: When Hammering Nails with a Microscope Makes Sense

O. Mezhenskyi, D. Kravchuk

Bogomoletz Institute of Physiology, National Academy of Sciences of Ukraine
olegmezh@biph.kiev.ua

Foundational models (FMs–large pre-trained neural architectures–are transforming modern biology by providing universal representations learned from massive, heterogeneous, and often unlabeled datasets. Unlike classical task-specific machine-learning models, FMs can be fine-tuned for genomics, cheminformatics, bioimaging, and physiological signal analysis with minimal amounts of labeled data. This mini-review summarizes key applications of DNABERT, MolBERT, DiffDock, and Segment Anything, highlighting their advantages in accuracy, generalizability, and multimodal integration. We also outline the potential of FMs in physiology and neurophysiology, where they may unify signals from patch-clamp recordings, microelectrode measurements, and calcium imaging into a single analytical framework..
Keywords : foundational models; machine learning; DNABERT; MolBERT; DiffDock; Segment Anything; bioinformatics; neurophysiology; patch-clamp; image analysis.

INTRODUCTION

Biological data in recent years has grown not only in volume but also in complexity, with the curse of dimensionality making it impossible to process everything properly. Modern technologies such as single-cell RNA sequencing, cryo-electron microscopy, spatial transcriptomics, and high-throughput chemical screening generate terabytes of heterogeneous data each day, in addition to already existing classic methods. One of the solutions to simplify the processing of such vast volumes of data includes computational methods. Traditionally, researchers have relied on task-specific machine learning (ML) models — convolutional neural networks for imaging, random forests for property prediction, or recurrent networks for DNA sequences [1]. While powerful within their niches, these models face several key limitations:

Dependence on labeled data: Most require thousands of labeled examples per task, which is costly and time-consuming to generate.

Lack of transferability: Models trained for one dataset or problem rarely generalize to another without extensive retraining.

Narrow focus: They often ignore broader biological context, limiting interpretability and cross-modal reasoning.

One of the ways to overcome those shortcomings is the use of foundational models (FMs). Rather than being trained for one purpose, FMs are pre-trained on massive, diverse, and often unlabeled datasets to learn universal representations of biological entities – whether sequences, structures, molecules, or images. Once pre-trained, they can be adapted to new tasks with minimal effort, often achieving top performance even when labeled data are scarce [2].

With FMs, the difference lies in their scaling behavior and emergent capabilities – as they grow larger and are exposed to broader data, they capture deeper biological principles, enabling zero-shot generalization and transfer across species or modalities. Another difference between FMs and classical ML is the number of parameters used for model operation. While classical models may use tens to hundreds

of thousands of parameters (weights), the smallest FMs start from millions and can scale up to several trillion parameters, making them extremely heavyweight and cumbersome to operate. Fortunately, this downside rarely matters, since fine-tuning usually affects only the last layers of the model – typically involving only thousands of parameters. This keeps the process of FM fine-tuning almost as fast as for classical models, although inference time is usually higher [2,3].

In this mini-review, we discuss specific applications of FMs in biology and how they can enhance data acquisition and analysis.

Biological Computer Vision: FaceID for cells

One of the most common challenges in modern biology is the automated analysis of microscopy images to identify or classify cells – a task central to drug discovery, histopathology, and basic research. Traditional CNN-based pipelines require task-specific training and often fail to generalize across imaging modalities, experimental setups, cell types, or even individual researchers [4,5].

The FM approach in this case is represented by the Segment Anything Model (SAM) [6], a foundational vision transformer originally developed for general image segmentation, which has been adapted for biological imaging with remarkable success. Pre-trained on billions of images (mostly from non-biological sources), SAM can be fine-tuned with minimal labeled data (hundreds to thousands of images) to segment cells, nuclei, or organelles across diverse imaging conditions. For example, a ready-to-use Cellpose-SAM model [7], fine-tuned specifically for cellular segmentation, was able to segment over 100 different cell types from more than 20 microscopy platforms with approximately 95% accuracy. Classical models would require a separate network for each condition to achieve comparable performance. Another benefit of SAM is its embeddings, which can serve as foundational representations of the input data [8]. Downstream models trained on these embeddings can classify specific cell states (e.g., stages of the cell cycle), detect rare phenotypes, or even predict transcriptomic signatures from morphology.

Summing up, the use of FMs in CV-based tasks is well justified, as it significantly accelerates analysis while maintaining excellent accuracy. CV-based use cases of FMs are not limited to microscopy – they can be extended to methods like fMRI and other imaging modalities, accurately identifying specific brain regions. One example includes NeuroSTORM [9], trained on more than 20 million fMRI images, which can be used for diagnostic and research-specific tasks such as cognitive phenotype classification.

Genomics and Transcriptomics: Reading the book of DNA between the lines

FMs can work not only in CV but also in other fields that benefit from computational pattern recognition. One such task is understanding how non-coding DNA regulates gene expression by finding gene-specific promoters or enhancer regions. Classical models such as CNN-based DeepSEA or logistic regression classifiers can predict enhancer or promoter activity but require carefully curated datasets and struggle with cross-species generalization [10].

A well-known FM in this field is DNABERT, which applies transformer architectures to DNA sequences by treating nucleotides as "words" and learning context-dependent representations of these "words." Pre-trained on billions of base pairs from annotated reference genomes, DNABERT captures patterns of motifs, chromatin context, and higher-order syntax without explicit supervision [11,12].

When fine-tuned for enhancer prediction, DNABERT outperforms classical models by up to 15% AUROC and can even generalize across species – accurately predicting enhancers in mouse using a model trained solely on human data. Furthermore, its embeddings can be reused for downstream tasks such as transcription factor binding prediction, variant effect prioritization, or CRISPR target scoring

with minimal retraining. Recent studies also show that FMs(Evo2 specifically) embeddings can be used to construct phylogenetic trees, capturing evolutionary relationships between species [13,14].

Chemical Property Prediction: Bringing SMILE to science

Predicting molecular properties such as solubility, toxicity, permeability, or binding affinity is critical in pharmaceutical research and has long been an area of computational focus. Classical Quantitative Structure Activity Relationship (QSAR) approaches rely on manually or semi-automatically engineered descriptors or task-specific graph neural networks (GNNs), which often fail in low-data regimes or when encountering structurally novel compounds [15,16].

The FM MolBERT [17,18], a BERT-style chemical language model, is pre-trained on tens of millions of SMILES strings (a compact textual representation of molecular structure, e.g., CCO for ethanol) using masked-token prediction. The model learns deep chemical semantics – capturing both local substructures and global molecular context – representing each molecule as a set of embeddings.

When fine-tuned for ADMET prediction on the MoleculeNet [19] benchmark dataset, MolBERT consistently outperforms classical GNNs and ML-QSAR methods, improving prediction accuracy by 5–20% while requiring an order of magnitude less labeled data.

The power of MolBERT lies in its transferability: a single pre-trained model can predict solubility, logP, blood-brain barrier permeability, and toxicity without retraining from scratch. It can even generalize to molecules outside its original training domain. Moreover, MolBERT embeddings enable zero-shot similarity search and few-shot virtual screening – capabilities unavailable to traditional QSAR models.

Molecular Structure and Interaction Prediction: making a difference with a DiffDock

Predicting how small molecules interact with proteins is fundamental to drug design, yet classical computational methods such as docking – which rely on physics-based scoring functions – are slow and often inaccurate for flexible ligands or novel binding sites.

The FM DiffDock [20], based on a diffusion architecture for generative modeling, learns the distribution of plausible protein–ligand complexes. Pre-trained on millions of binding poses from the PDBBind database [21], DiffDock can generate likely docking conformations in a single forward pass — making inference far faster than classical docking and several orders of magnitude faster than molecular dynamics simulations. On the PoseBusters benchmark [22], DiffDock achieves a top-1 success rate (within 2 A RMSD deviation).

Biophysics and physiology: clamping the patch

Physiological regulation arises from a symphony of interacting systems – neural, endocrine, and metabolic. Historically, each has been analyzed in isolation simply because joint modeling was computationally unfeasible. Foundational models, however, are built to handle multimodal data and could finally bring these signals together.

A future physiological FM could be trained on diverse recordings – from microelectrodes, patch-clamp traces, or calcium-imaging data – to detect and classify events like neuronal firing or ion-channel gating. Going further, it could perform generative forecasting, predicting how unseen physiological events might unfold under different conditions.

Although such a model doesn't yet exist publicly, benchmarks like MIMIC-III and the PhysioNet Challenges provide ready frameworks for developing and evaluating it [23,24].

Summary: Sometimes Bigger Really Is Better

Classical ML models are elegant and efficient, but usually excel only within narrow boundaries. Foundational models, by contrast, are heavy, expensive, and unwieldy – yet they often brute-force their way across tasks and modalities with unmatched generality.

A single pre-trained FM can now anchor dozens of downstream applications, cutting data requirements, improving interpretability, and bridging previously separate biological domains.

So far, FMs dominate where data naturally come as images, sequences, or tables. But extending them into biophysics – from ion-channel recordings to whole-organ physiology – could open an entirely new frontier. Sometimes, as it turns out, hammering nails with a microscope is exactly the right kind of madness.

О. Меженський, Д. Кравчук

ЗАСАДНИЧІ МОДЕЛІ В БІОЛОГІЇ: КОЛИ ЗАБИВАТИ ЦВЯХИ МІКРОСКОПОМ – ЦІЛКОМ СЛУШНА ІДЕЯ

Інститут фізіології ім. О. О. Богомольця НАН України
olegmezh@biph.kiev.ua

Засадничі моделі (Foundational Models, FMs) — це великі попередньо натреновані нейронні архітектури, які формують універсальні представлення, навчені на масивних, гетерогенних та часто неанотованих біологічних даних. На відміну від класичних моделей машинного навчання, що орієнтовані на виконання вузьких завдань, FMs можна ефективно донавчати для аналізу даних у геноміці, хемоінформатиці, біологічній візуалізації та фізіології навіть за умов обмеженої кількості розмічених прикладів. У цьому короткому огляді розглянуто ключові застосування моделей DNABERT, MolBERT, DiffDock і Segment Anything, підкреслено їхню вищу точність, універсальність та здатність до інтеграції кількох типів біологічних даних. Окремо обговорюються перспективи застосування засадничих моделей у фізіології та нейрофізіології, де вони можуть об'єднувати сигнали петч-клемпу, мікроелектродних та кальцієвих записів у єдину аналітичну платформу.

Ключові слова: засадничі моделі; машинне навчання; DNABERT; MolBERT; DiffDock; Segment Anything; біоінформатика; нейрофізіологія; петч-клемп; аналіз зображень.

REFERENCES

1. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol. 2022;23(1):40–55.
2. Awais M, Naseer M, Khan S, Anwer RM, Cholakkal H, Shah M, et al. Foundational models defining a new era in vision: A survey and outlook. IEEE Trans Pattern Anal Mach Intell. 2023;47(4):2245–64.
3. Neidlinger P, Nahhas OSME, Muti HS, Lenz T, Hoffmeister M, Brenner H, et al. Benchmarking foundation models as feature extractors for weakly-supervised computational pathology. arXiv. 2024. Available from: http://arxiv.org/abs/2408.15823.
4. Tran DH, Meunier M, Cheriet F. Multi-domain learning CNN model for microscopy image classification. arXiv. 2023. Available from: http://arxiv.org/abs/2304.10616
5. Danuser G. Computer Vision in Cell Biology. Cell. 2011;147(5):973–8.
6. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. Segment Anything. Proc IEEE Int Conf Comput Vis. 2023;3992–4003.
7. Stringer C, Wang T, Michaelos M, Pachitariu M. Cellpose: a generalist algorithm for cellular segmentation. Nat Methods. 2021;18(1):100–6.
8. Low-dimensional embeddings of high-dimensional data. arXiv. 2025. Available from: https://arxiv.org/html/2508.15929v1.
9. Wang C, Jiang Y, Peng Z, Li C, Bang C, Zhao L, et al. Towards a general-purpose foundation model for fMRI analysis. 2025.
10. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015;12(10):931–4.
11. Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics. 2021;37(15):2112–20.
12. Zhou Z, Ji Y, Li W, Ramana D, Davuluri V, et al. DNABERT-2: Efficient foundation model and benchmark for multi-species genome. 2024. Available from: https://arxiv.org/pdf/2306.15006v2.
13. Finding the Tree of Life in Evo2. 2025. Available from: https://www.goodfire.ai/research/phylogeny-manifold.
14. Brixi G, Durrant MG, Ku J, Poli M, Brockman G, Chang D, et al. Genome modeling and design across all domains of life with Evo2. bioRxiv. 2025. Available from: https://www.biorxiv.org/content/10.1101/2025.02.18.638918v1
15. Keyvanpour MR, Shirzad MB. An analysis of QSAR research based on machine learning concepts. Curr Drug Discov Technol. 2021;18(1):17–30.
16. Graph neural networks for materials science and chemistry. Commun Mater. 2022. Available from: https://www.nature.com/articles/s43246-022-00315-6.
17. Chithrananda S, Grand G, Ramsundar B. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. 2020. Available from: https://arxiv.org/pdf/2010.09885.
18. Li J, Jiang X. Mol-BERT: An effective molecular representation with BERT for molecular property prediction. Wirel Commun Mob Comput. 2021; 2021:7181815.
19. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: A benchmark for molecular machine learning. arXiv. 2018. Available from: http://arxiv.org/abs/1703.00564
20. Corso G, Stärk H, Jing B, Barzilay R, Jaakkola T. DiffDock: Diffusion steps, twists, and turns for molecular docking. ICLR 2023. Available from: https://arxiv.org/pdf/2210.01776

21. Wang R, Fang X, Lu Y, Yang CY, Wang S. The PDBbind database: methodologies and updates. J Med Chem. 2005;48(12):4111–9.
22. Buttenschoen M, Morris GM, Deane CM. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. Chem Sci. 2023;15(9):3130–9.
23. Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035.
24. Clifford GD, Liu C, Moody B, Lehman LH, Silva I, Li Q, et al. AF classification from a short single-lead ECG recording: The PhysioNet/Computing in Cardiology Challenge 2017. Comput Cardiol. 2017;44:1–4.