ChimeraLM: A genomic language model for detecting whole genome amplification artifacts in single-cell sequencing

Yangyang Li[1], Qingxiang Guo[1†], Rendong Yang[1,2*]

[1]Department of Urology, Northwestern University Feinberg School of Medicine, 303 E Superior St, Chicago, 60611, IL, USA.
[2]Robert H. Lurie Comprehensive Cancer Center, Northwestern University Feinberg School of Medicine, 675 N St Clair St, Chicago, 60611, IL, USA.

*Corresponding author(s). E-mail(s): rendong.yang@northwestern.edu;
Contributing authors: yangyang.li@northwestern.edu;
qingxiang.guo@northwestern.edu;
†These authors contributed equally to this work.

Abstract

this is a abtract

Keywords: Whole Genomics Amplification, Genomic Language Model

Main

Single-cell genomics has revolutionized our understanding of cellular heterogeneity and development by enabling the characterization of individual cells rather than bulk populations [1, 2]. This approach has proven instrumental in uncovering rare cell types, tracking developmental trajectories, and identifying somatic mutations that drive disease progression. However, the limited amount of DNA present in a single cell, typically only a few picograms, poses significant technical challenges for comprehensive genomic analysis [3, 4].

To overcome this fundamental limitation, Whole Genome Amplification (WGA) has become an essential preprocessing step in single-cell genomic studies [5, 6]. Various

WGA techniques, including Multiple Displacement Amplification (MDA), Multiple Annealing and Looping-based Amplification Cycles (MALBAC), and other emerging methods, can amplify the entire genome from a single cell by several orders of magnitude, generating sufficient DNA material for high-coverage sequencing [7–10]. This amplification enables researchers to achieve the depth and breadth of coverage necessary for reliable variant calling, copy number analysis, and structural variation detection.

Despite its critical role in single-cell genomics, WGA introduces systematic artifacts that can significantly impact downstream analyses [11, 12]. Among the most problematic are chimeric sequences—artificial DNA constructs formed when DNA fragments from different genomic loci are erroneously joined during the amplification process [10–12]. These chimeric artifacts can manifest as false-positive structural variations that do not exist in the original cell [11]. The presence of such artifacts poses a substantial challenge for accurate Structural Variation (SV) detection, potentially leading to misinterpretation of genomic rearrangements and their biological significance.

Current computational approaches for identifying WGA-induced artifacts rely primarily on coverage-based metrics and read-pair orientation patterns [12, 13]. However, these methods often fail to distinguish between genuine structural variations and amplification artifacts, particularly when chimeric sequences exhibit complex rearrangement patterns or occur in repetitive genomic regions [14, 15]. The lack of robust artifact detection methods has limited the reliability of structural variant analysis in single-cell studies and hindered the full realization of single-cell genomics' potential.

To address these challenges, we developed ChimeraLM, a genomic language model specifically designed to detect chimeric artifacts introduced by whole genome amplification. By leveraging deep learning approaches to capture sequence patterns and contextual information in genomic reads [16–18], ChimeraLM can effectively distinguish between genuine biological sequences and WGA-induced chimeric artifacts. This approach represents a significant advancement in single-cell genomic analysis, offering improved accuracy in artifact detection and enabling more reliable structural variant analysis in single-cell studies. This methodology represents a significant advancement in single-cell genomic analysis, offering a principled approach to improve the reliability of structural variant detection and enable more precise characterization of genomic alterations in individual cells.

In this study, we present ChimeraLM, demonstrate its superior performance compared to existing methods, and illustrate its practical applications in genomic studies.

Results

ChimeraLM integrates seamlessly into single-cell genomic workflows

To systematically address WGA-induced chimeric artifacts, we developed ChimeraLM as an integrated component of single-cell genomic analysis pipelines (Figure 1a). Our approach leverages the standard single-cell workflow, beginning with cellular isolation through FACS or microfluidics-based sorting, followed by DNA extraction and whole

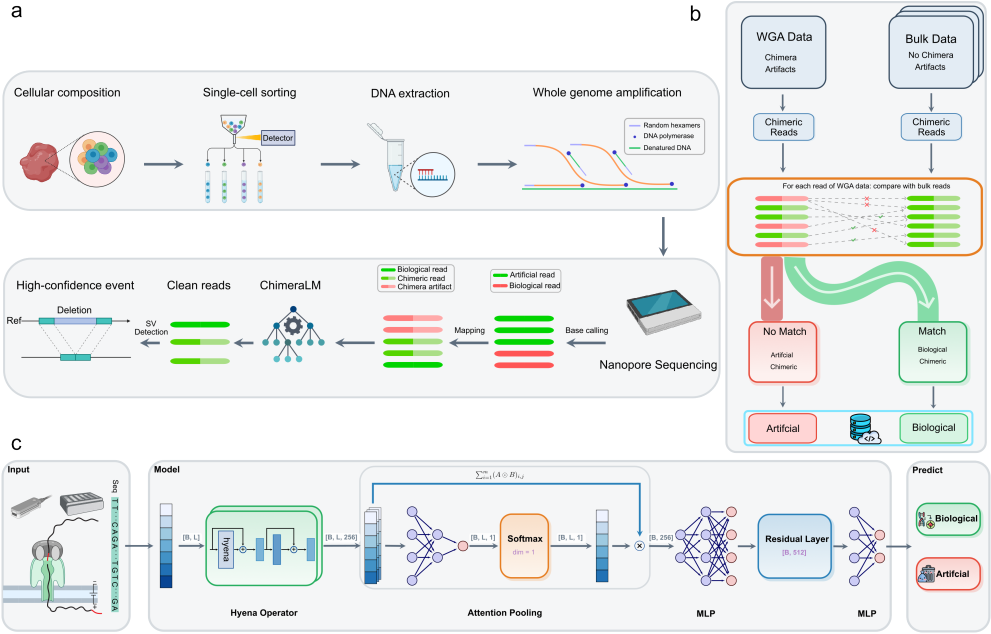

Fig. 1 ChimeraLM workflow and architecture for detecting whole genome amplification artifacts in single-cell sequencing. (a) Integration of ChimeraLM into single-cell genomic analysis workflows. Single cells are isolated from heterogeneous cellular populations through sorting technologies, followed by DNA extraction and WGA to generate sufficient material for sequencing. WGA introduces chimeric artifacts through random hexamers, DNA polymerase extension, and denatured DNA template switching. Sequencing reads from WGA-amplified samples contain both biological reads (green) and chimeric artifacts (red). ChimeraLM processes these reads to classify them as biological or artificial, enabling clean reads to proceed to structural variant analysis for high-confidence event detection such as deletions. (b) Dataset construction strategy for supervised learning. Training data is generated by comparing WGA sequencing reads against matched bulk sequencing data from the same biological sample. Bulk data contains only genuine biological sequences (no chimeric artifacts), while WGA data contains both biological reads and chimeric artifacts. Each WGA read is aligned against bulk data: reads that successfully match are labeled as "biological" (green pathway), while reads that fail to match are labeled as "artificial chimeric" (red pathway). This comparative approach provides reliable ground truth labels for training ChimeraLM in a supervised learning framework. (c) ChimeraLM neural network architecture. Input DNA sequences are tokenized and processed through a deep learning pipeline optimized for genomic sequence analysis. The architecture employs Hyena operators for efficient long-range dependency modeling, followed by attention pooling to aggregate variable-length sequence features. multilayer perceptron (MLP) components with residual connections process the pooled features to learn complex patterns distinguishing biological sequences from chimeric artifacts. The final output layer produces binary classification probabilities, predicting whether each input sequence represents a biological read or an artificial chimeric artifact.

genome amplification using established protocols. Amplified genomic material is then processed through long-read sequencing platforms such as Nanopore technology to generate comprehensive genomic coverage.

ChimeraLM operates at a critical juncture in the analysis pipeline, positioned between initial read processing and downstream analysis, for example, structural variants detection (Figure 1a). Following standard quality filtering and read cleaning procedures, ChimeraLM evaluates each sequencing read to classify it as either biological or chimeric artifacts. This binary classification enables the selective retention of authentic genomic sequences while filtering out amplification artifacts before they can impact downstream analyses.

The filtered, high-quality biological reads are subsequently processed through conventional structural variant detection algorithms, enabling the identification of genuine genomic alterations such as deletions, duplications, and other rearrangements. By removing chimeric sequences upstream of variant calling, ChimeraLM ensures that detected structural variants represent true biological events rather than technical artifacts introduced during amplification (Figure 1a).

This workflow design allows ChimeraLM to integrate with existing single-cell genomic pipelines without requiring substantial modifications to established protocols. The method provides a versatile solution for improving the accuracy of single-cell genomic studies across diverse research applications.

Training dataset construction enables supervised learning of chimeric patterns

To train ChimeraLM for accurate chimeric artifact detection, we developed a novel dataset construction strategy that leverages paired WGA and bulk sequencing data from the same biological samples (Figure 1b). This approach exploits the fundamental difference between WGA and bulk sequencing: while WGA data contains both biological reads and chimeric artifacts introduced during amplification, bulk sequencing data from the same sample contains only genuine biological sequences.

Our ground truth labeling strategy compares each chimeric read of WGA data against the corresponding one of bulk sequencing dataset (Figure 1b). Chimeric reads that successfully matched to bulk data with high confidence are classified as "biological" indicating they represent authentic genomic sequences present in the original sample. Conversely, reads that fail to match bulk sequences are labeled as "artificial chimeric" artifacts, as they represent artificial constructs generated during the WGA process rather than genuine genomic content (Figure 1b).

This comparative approach generates a comprehensive labeled dataset where each chimeric read of WGA receives a binary classification based on its presence or absence in the matched bulk control. The resulting dataset captures the full spectrum of chimeric artifacts naturally occurring during WGA while providing reliable ground truth labels for model training (Figure 1b).

Following dataset construction, we partitioned the labeled reads into training, validation, and test sets to ensure robust model development and unbiased performance evaluation. The training set was used for model parameter optimization, the validation set for hyperparameter tuning and model selection, and the test set was reserved for final performance assessment. This rigorous data splitting strategy ensures that ChimeraLM's reported performance metrics reflect its ability to generalize to previously unseen WGA data.

ChimeraLM architecture leverages modern genomic language modeling advances

ChimeraLM employs a sophisticated neural architecture specifically designed for genomic sequence analysis and chimeric artifact detection (Figure 1 c). The model enables single-base pair resolution and accepts DNA sequences as input, which are first tokenized and encoded into numerical representations suitable for deep learning processing. This encoding preserves the sequential nature of genomic information while enabling efficient computation on modern hardware.

The core of ChimeraLM's architecture consists of Hyena operators, a recent advancement in sequence modeling that provides computational advantages over traditional transformer attention mechanisms while maintaining the ability to capture long-range dependencies in genomic sequences. Hyena operators enable ChimeraLM to process variable-length sequencing reads efficiently while learning complex patterns that distinguish biological sequences from chimeric artifacts. Moreover, the hyena operators are initiallized with HyenaDNA [18], a pre-trained long-context genomic language model, which provides a strong foundation for learning genomic sequence features relevant to chimeric detection.

Following the Hyena operator layers, ChimeraLM incorporates an attention pooling mechanism that aggregates sequence-level features. This pooling strategy allows the model to handle reads of varying lengths while focusing computational attention on the most informative regions for chimeric detection. The attention weights learned during training provide interpretability regarding which sequence features contribute most strongly to classification decisions.

The aggregated features are then processed through multiple MLP components arranged in a residual architecture. This design enables gradient flow optimization during training while allowing the model to learn both low-level sequence motifs and high-level compositional patterns indicative of chimeric artifacts. The residual connections help prevent vanishing gradients and improve model convergence during training on large genomic datasets.

The final output layer produces a binary classification predicting whether each input sequence represents a biological read or an artificial chimeric artifact. This end-to-end architecture enables ChimeraLM to learn directly from raw sequence data without requiring manual feature engineering, allowing the model to discover complex patterns that may not be apparent through traditional bioinformatics approaches.

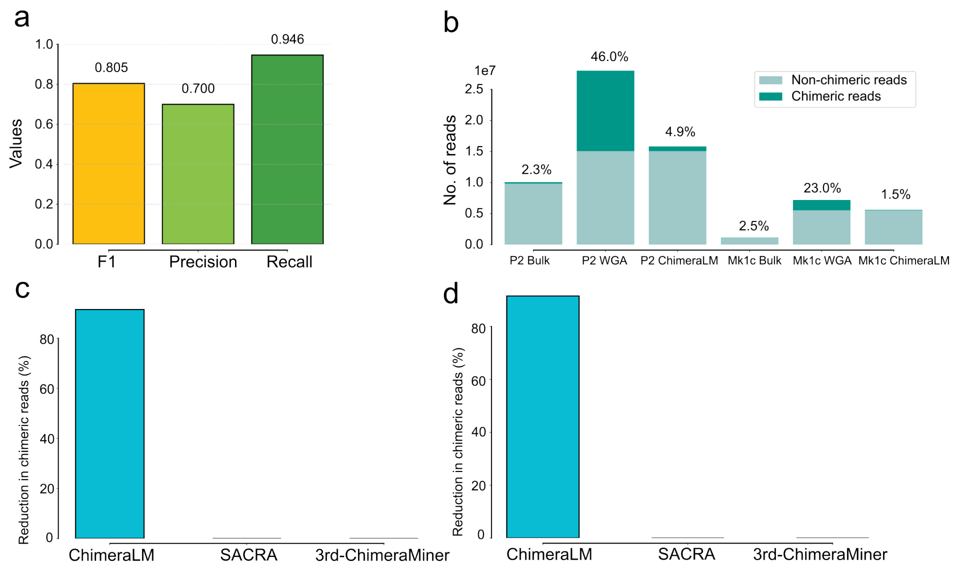

Fig. 2 Problem and Model

Methods

MDA sequencing

Train data construction

Model architecture

Model training

SV evaluation

Supplementary information. This separation aligns with how many transcript assembly algorithms work:

1. First, chains of exons and splice junctions are identified from the data
2. Then, potential transcripts are derived by traversing the graph in different ways
3. Finally, relationships between different transcript graphs are established

Acknowledgements. Acknowledgements are not compulsory. Where included they should be brief. Grant or contribution numbers may be acknowledged.

Please refer to Journal-level guidance for any specific requirements.

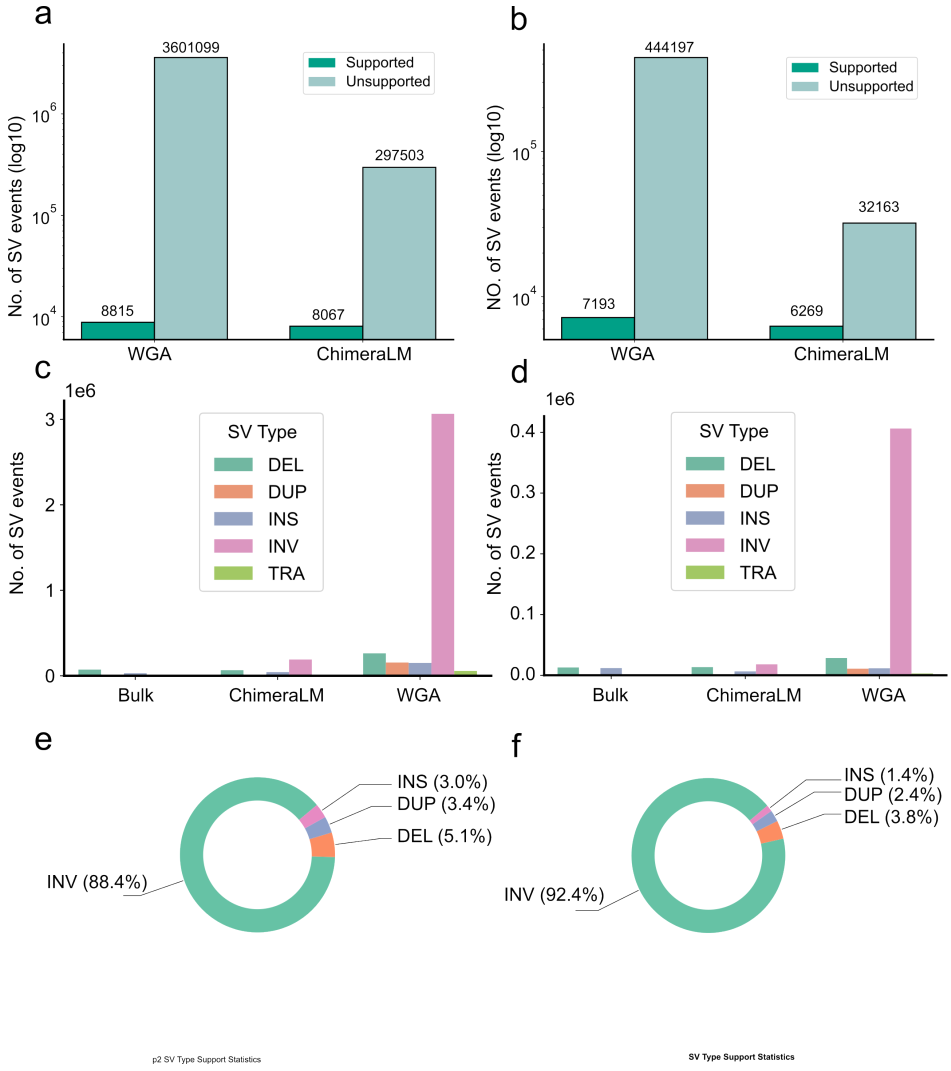

Fig. 3 Problem and Model

Declarations

Some journals require declarations to be submitted in a standardised format. Please check the Instructions for Authors of the journal to which you are submitting to see if you need to complete this section. If yes, your manuscript must contain the following sections under the heading 'Declarations':

- Funding
- Conflict of interest/Competing interests (check journal-specific guidelines for which heading to use)

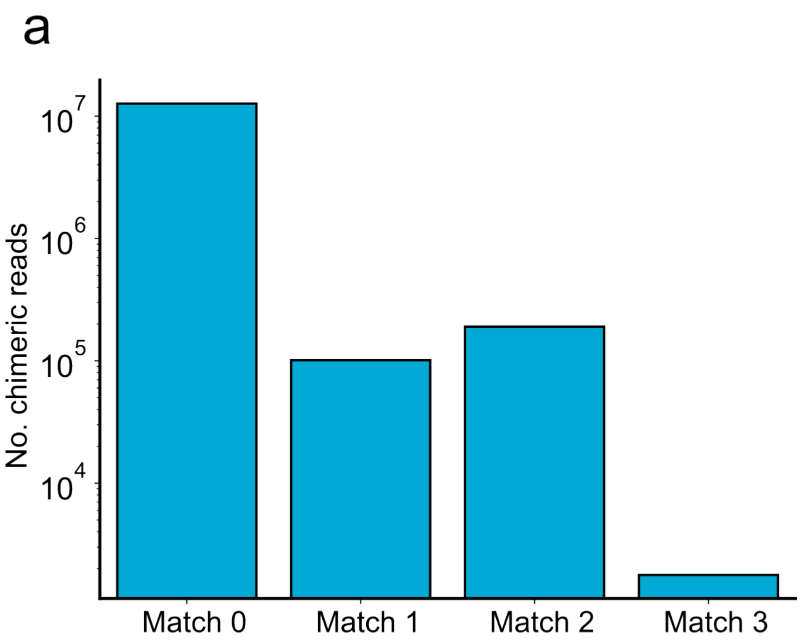

Fig. 4 Problem and Model

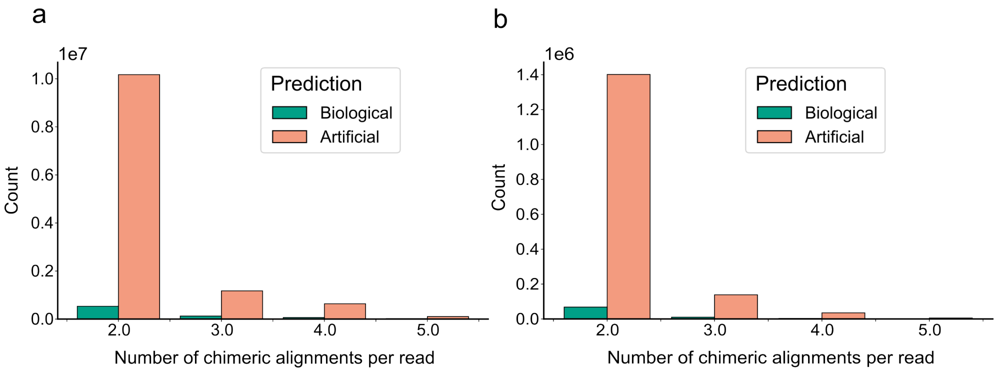

Fig. 5 Problem and Model

- Ethics approval and consent to participate
- Consent for publication
- Data availability
- Materials availability
- Code availability
- Author contribution

If any of the sections are not relevant to your manuscript, please include the heading and write 'Not applicable' for that section.

Editorial Policies for:

Springer journals and proceedings: https://www.springer.com/gp/editorial-policies

Nature Portfolio journals: https://www.nature.com/nature-research/editorial-policies

Scientific Reports: https://www.nature.com/srep/journal-policies/editorial-policies

BMC journals: https://www.biomedcentral.com/getpublished/editorial-policies

Acronyms

MALBAC Multiple Annealing and Looping-based Amplification Cycles 2
MDA Multiple Displacement Amplification 2
MLP multilayer perceptron 3, 5

SV Structural Variation 2

WGA Whole Genome Amplification 1–4

Appendix A Section title of first appendix

An appendix contains supplementary information that is not an essential part of the text itself but which may be helpful in providing a more comprehensive understanding of the research problem or it is information that is too cumbersome to be included in the body of the paper.

References

[1] Kalef-Ezra, E. *et al.* Single-cell somatic copy number variants in brain using different amplification methods and reference genomes. *Communications Biology* **7**, 1288 (2024).

[2] Sun, C. *et al.* Mapping recurrent mosaic copy number variation in human neurons. *Nature communications* **15**, 4220 (2024).

[3] Leung, M. L. *et al.* Highly multiplexed targeted dna sequencing from single nuclei. *Nature protocols* **11**, 214–235 (2016).

[4] Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics* **17**, 175–188 (2016).

[5] Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622–1626 (2012).

[6] Huang, L., Ma, F., Chapman, A., Lu, S. & Xie, X. S. Single-cell whole-genome amplification and sequencing: methodology and applications. *Annual review of genomics and human genetics* **16**, 79–102 (2015).

[7] De Bourcy, C. F. *et al.* A quantitative comparison of single-cell whole genome amplification methods. *PloS one* **9**, e105585 (2014).

[8] Biezuner, T. *et al.* Comparison of seven single cell whole genome amplification commercial kits using targeted sequencing. *Scientific reports* **11**, 17171 (2021).

[9] Fu, Y. *et al.* Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification. *Proceedings of the National Academy of Sciences* **112**, 11923–11928 (2015).

[10] Agyabeng-Dadzie, F. *et al.* Evaluating the benefits and limits of multiple displacement amplification with whole-genome oxford nanopore sequencing. *Molecular Ecology Resources* e14094 (2025).

[11] Lu, N., Qiao, Y., Lu, Z. & Tu, J. Chimera: The spoiler in multiple displacement amplification. *Computational and Structural Biotechnology Journal* **21**, 1688–1696 (2023).

[12] Lu, N. *et al.* Exploration of whole genome amplification generated chimeric sequences in long-read sequencing data. *Briefings in Bioinformatics* **24**, bbad275 (2023).

[13] Kiguchi, Y., Nishijima, S., Kumar, N., Hattori, M. & Suda, W. Long-read metagenomics of multiple displacement amplified dna of low-biomass human gut phageomes by sacra pre-processing chimeric reads. *DNA Research* **28**, dsab019 (2021).

[14] Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome biology* **20**, 117 (2019).

[15] Mahmoud, M. *et al.* Structural variant calling: the long and the short of it. *Genome biology* **20**, 246 (2019).

[16] Dalla-Torre, H. *et al.* Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods* **22**, 287–297 (2025).

[17] Zhou, Z. *et al.* Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006* (2023).

[18] Nguyen, E. *et al.* Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems* **36**, 43177–43201 (2023).