Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Genetics & Development

Review

Deep learning for psychiatric genomics: from tools to applications

Junhao Liu, Siwei Xu, Dongbo Sun, Chaoyang Wang and Jing Zhang

The genetic architecture of psychiatric disorders is highly complex, with genome-wide association studies implicating thousands of risk loci. A central challenge is that most of these variants are located in noncoding regions, making it difficult to elucidate their regulatory consequences within the brain's intricate cellular landscape. The recent convergence of advanced artificial intelligence, particularly deep learning (DL), has catalyzed a paradigm shift by providing powerful tools to address this gap. This review traces the evolution of DL in genomics, beginning with task-specific models. We then examine the transformative impact of foundation models, pretrained neural networks that learn the 'language' of biology, including genomic language models, single-cell foundation models, and large language models originally trained on natural language. Finally, we survey applications to key problems in psychiatric genomics. We hope this review provides a comprehensive overview of recent advances in DL for genomics and serves as a bridge to help researchers in psychiatric genomics more effectively understand and apply these frontier methods to guide the development of novel therapeutic strategies for psychiatric disorders.

Address
Department of Computer Science, University of California, Irvine, 6210 Donald Bren Hall, Irvine, 92697 CA, USA

Corresponding author: Zhang, Jing (zhang.jing@uci.edu)

Current Opinion in Genetics & Development 2026, **97**:102442

This review comes from a themed issue on **Molecular and Genetic Basis of Disease**

Edited by **Hyejung Won** and **Michael Ziller**

For complete overview of the section, please refer to the article collection, "Molecular and Genetic Basis of Disease (2026)"

Available online 14 February 2026

https://doi.org/10.1016/j.gde.2026.102442

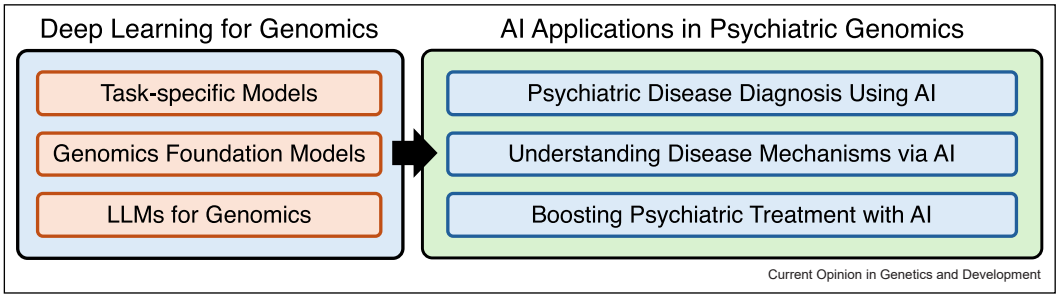
Introduction

Genomics has emerged as a transformative modality in psychiatric research, imparting an objective biological framework for disorders historically characterized by subjective symptomatology [1]. This approach elucidates the molecular etiology of disease by pinpointing genetic risk variants and the functional pathways they impair, thereby revealing novel therapeutic targets [2]. A key advantage lies in genomics' capacity to quantify disease susceptibility during the pre-symptomatic phase, avenues for preemptive intervention, and primary prevention [3]. Concurrently, rapid technological innovations, coupled with consortium-based initiatives promoting open data sharing, have catalyzed an exponential expansion of large-scale, multi-omic datasets [4]. This burgeoning resource provides an unprecedented foundation for deconvoluting the pronounced cellular and phenotypic heterogeneity inherent to psychiatric disorders. Motivated by these developments, we herein present a systematic evaluation of the burgeoning role of artificial intelligence (AI) in advancing this field.

Genomic research has progressively mapped the genetic architecture of human disease, enabling clinical translations from targeted oncology to early diagnostics [5]. While large-scale biobanks now leverage these insights for predictive medicine, the past decade has witnessed a particular renaissance in psychiatric genomics, driven by collaborative consortia and advanced analytical tools [6–8]. These efforts reveal that the field's complex and context-dependent central challenges, including a highly polygenic architecture, significant phenotypic heterogeneity, the inaccessibility of brain tissue, and pervasive gene–environment interactions [9]. Consequently, the application of AI to this domain remains a formidable yet essential frontier for unraveling the biological underpinnings of psychiatric illness.

The profound complexity of psychiatric genomics renders advanced computational approaches, particularly AI and deep learning (DL) [10], indispensable for deciphering high-dimensional data. These methods excel at identifying subtle, non-additive patterns, such as epistasis and gene–environment interactions, that elude conventional statistics. [11] Accordingly, this review synthesizes recent DL advancements tailored for genomic data and their application to psychiatric disorders. We survey task-specific architectures before examining foundation models, which use pretraining to capture fundamental biological 'language' from vast molecular datasets. Finally, we highlight how these tools

Figure 1



Conceptual framework of the review.

are being deployed to unravel psychiatric disease mechanisms, outlining a path from computational innovation to therapeutic insight. Figure 1 summarizes the key concepts discussed in this review. This overview aims to bridge disciplines, equipping researchers with the conceptual framework needed to harness these powerful methods.

Deep learning methods for fundamental genomics

This section reviews DL approaches for genomics, spanning task-specific neural architectures to foundation models pretrained via self-supervised learning. Key methods discussed are summarized in Table 1.

Task-specific deep learning methods on genomics

DL enables flexible model design and learns task-relevant representations directly from data, reducing reliance on manual feature engineering. Standard architectures, including convolutional neural networks (CNNs) [12], Recurrent neural networks [13], Transformers [14], and their recent variants [15,16], offer complementary strengths, and selecting or combining them based on task demands and computational constraints typically leads to improved performance.

A primary application of task-specific DL methods in psychiatric genomics is interpreting noncoding variation through models that predict how sequence alterations disrupt the cis-regulatory code. Several architectures exemplify this approach: DeepSEA [17], for instance, employs a CNN to forecast chromatin effects, while ChromBPNet [18] explicitly deconvolves technical biases to model base-resolution accessibility. Further advancing this capability, the Borzoi model [19] utilizes a hybrid Transformer–CNN architecture to directly predict tissue-specific RNA-seq coverage from sequence alone. Expanding beyond these core tasks, subsequent

Table 1

Summary of DL methods discussed in this review.

Section	Method	Architecture	Input	Output
2.1	DeepSEA	CNN	DNA sequences	Chromatin effects of noncoding variants
2.1	Enformer	Transformer	Long-range DNA sequences	Gene expression
2.1	Sei	CNN	DNA sequences	Chromatin regulatory activity
2.1	PromoterAI	Transformer variant	Promoter DNA sequences	Expression-altering effects of variants
2.1	Borzoi	Transformer + CNN	DNA sequences	RNA-seq coverage and splicing events
2.1	ChromBPNet	CNN	DNA sequences	Base-resolution chromatin accessibility
2.1	AlphaGenome	Transformer + CNN	DNA sequences	Regulatory variant effects
2.2.1	DNABERT	Transformer	DNA k-mers	Multi-species genome embeddings
2.2.1	GENA-LM	Transformer	Long genomic sequences	Genome embeddings for downstream tasks
2.2.1	Nucleotide Transformer	Transformer	DNA sequence (diverse genomes)	Versatile DNA sequence prediction
2.2.1	HyenaDNA	Hyena	Ultralong genomic sequences	Single-nucleotide resolution modeling
2.2.1	Caduceus	Mamba	Long genomic sequences	Bi-directional, reverse-complement embeddings
2.2.1	Evo	StripedHyena	DNA sequence (single-nucleotide)	Genome generation and prediction
2.2.1	Cell2Sentence	LLM	Cells represented as sentences	Cell generation and cell type prediction
2.2.1	GenePT	LLM	Literature-informed text embeddings	Gene functions and interactions
2.2.2	Geneformer	Transformer	Single-cell transcriptomes	Gene network dynamics and targets
2.2.2	scGPT	Transformer	Single-cell multi-omics	Cell states and perturbation responses
2.2.2	scMulan	Transformer	Single-cell data	Generated transcriptomes and cell identity
2.2.2	scFoundation	Transformer	Single-cell transcriptomes	Cell types and drug sensitivities
2.2.2	CellFM	Transformer	Large-scale transcriptomes	Cell types and regulatory networks

models have broadened the predictive scope to encompass specific elements such as promoter activity (PromoterAI) [20], deliver genome-wide functional annotations (Sei) [21], and capture long-range enhancer–promoter interactions (Enformer) [22]. This progression culminates in integrated frameworks such as AlphaGenome [23], which aim to unify diverse functional readouts within a single predictive model. Collectively, these methods have become indispensable tools for elucidating the molecular mechanisms of a broad spectrum of human diseases.

Genomics foundation models

In contrast to task-specific models, genomics foundation models acquire generalizable patterns through self-supervised learning on large, unlabeled datasets. We examine two foundational approaches: DNA language models and those based on transcriptomic data.

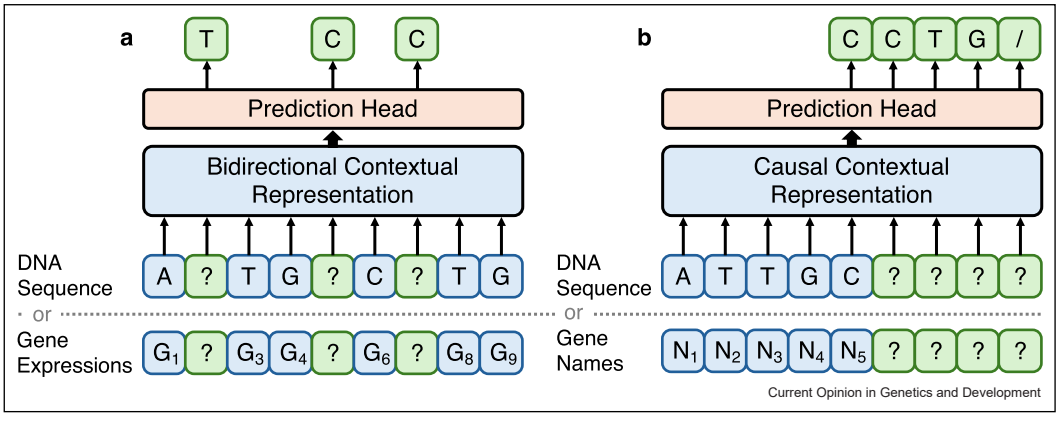
Genomics language models and their applications in genomics

In recent years, inspired by the remarkable success of large language models (LLMs), researchers have sought to explore similar advances in modeling DNA sequences. To enable self-supervised pretraining on genomic data, two pretraining objectives originally proposed in the natural language processing community have been adopted: masked token prediction (MTP) and next token prediction (NTP). An illustration of these objectives is shown in Figure 2. A representative series of models, including DNABERT [24,25] and GENA-LM [26], adopt the MTP objective with transformer-based architectures. The most recent version, DNABERT-2 [25], employs byte-pair encoding for DNA sequence tokenization and is pretrained on 32.49 billion base pairs, spanning both the human genome and genomes from 135 additional species. The Nucleotide Transformer [27] is another genomics language model leveraging the MTP objective, scaling up to 2.5 billion parameters.

Although transformers are highly effective, their quadratic computational complexity limits the context length for DNA sequences, making single-base-resolution modeling inefficient. To overcome this, several genomics language models based on linear attention mechanisms have been proposed. Key examples include HyenaDNA [15], Caduceus [28], and Evo [29,30]. HyenaDNA and Evo use the Hyena [15] and StripedHyena architectures [29], respectively, trained with the NTP objective, while Caduceus is based on the Mamba architecture [16] and employs MTP. All three models were pretrained on the human genome.

LLMs such as ChatGPT [31] have demonstrated strong capabilities across a wide range of human tasks, motivating researchers to investigate whether these models can also be applied to genomics. A key advantage of using existing LLMs is that they are typically pretrained on large-scale text corpora, including scientific literature, which provides them with latent knowledge of biological concepts. This raises the possibility that they can interpret genomic data without additional fine-tuning or pretraining on domain-specific datasets. For example, Hou and Ji [32] demonstrated that GPT-4 can produce accurate cell-type annotations. GenePT [33] further illustrates this idea by prompting LLMs with gene descriptions to generate informative gene embeddings that can then support downstream analyses such as cell-type annotation and batch-effect correction. Similarly, Cell2Sentence [34] reformulates downstream genomics tasks into a question-answering framework, enabling LLMs to be fine-tuned for improved performance on genomics-related queries.

Figure 2



Summary of pretraining strategies for foundation models across different input data types. **(a)** MTP is applied to DNA sequences or gene expression data, depending on the type of foundation model. **(b)** NTP is applied to DNA sequences or gene names.

Transcriptomic foundation models and their applications in genomics

In addition to gLMs, researchers have also developed foundation models for other genomic modalities, with single-cell RNA-seq data serving as a prominent example. Similar to DNA sequence foundation models, these approaches pretrain models on large-scale single-cell RNA-seq datasets, such as the Human Cell Atlas [35]. Representative works include Geneformer [36], scGPT [37], scMulan [38], scFoundation [39], and CellFM [40]. These models generally adopt MTP as their pretraining objective and employ various strategies to capture gene–gene interactions from the unordered gene expression vector of each cell. Evidence suggests that such cell-level foundation models are effective in tasks such as cell-type annotation and batch-effect correction. [37,40].

Constructing a pipeline to understand psychiatric genomics with artificial intelligence

The growing arsenal of genomic AI tools is being applied to tackle psychiatric diseases, aiming to improve diagnosis, treatment, and prevention through a genetic lens. It is important to note that while the DL approaches discussed herein demonstrate significant potential, they represent frontier research and currently remain at the proof-of-concept stage. A summary of the applications is presented in Table 2.

Psychiatric disease diagnosis using artificial intelligence

Disease diagnosis seeks to anticipate disease before symptom onset by integrating genetic and biological context. Psychiatric disorders are highly polygenic and heterogeneous, making single-nucleotide polymorphism (SNP)-based modeling alone insufficient. Sequence-informed approaches evaluate variant effects by substituting reference with alternative alleles in the native sequence context, yielding importance scores that capture long-range regulatory influences on enhancers, promoters, and gene expression, offering a more biologically nuanced view than conventional polygenic risk scores [19].

A central application of these sequence-based models, such as DeepSEA [17], ChromBPNet [18], and Enformer [22], is functional fine-mapping. GWAS identify disease-associated loci, but linkage disequilibrium often leaves hundreds of correlated variants per region, only a small fraction of which are truly causal [54]. By predicting the functional impact of each variant, sequence models can prioritize those most likely to be causal, guiding experimental follow-up.

Table 2

Summary of AI methods and applications in psychiatric genomics.

Category	Methods and models	Key applications and functions
Psychiatric disease diagnosis using AI	DeepSEA [17], ChromBPNet [18], Enformer [22]	Sequence-based modeling to prioritize causal noncoding variants and perform functional fine-mapping (e.g. in ASD).
	Transformer-based DNAm model [41]	Predicting DNAm levels at CpG sites from DNA sequence to map regulatory variants.
	Multimodal Frameworks [7,42]	Integrating scQTL, bulk eQTL, and GRNs to link genotypes to phenotypes and refine risk estimates.
	Clinical Integration Models [43,44]	Incorporating EHR, clinical variables, and wearable data for comprehensive risk profiling.
Understanding disease mechanisms via AI	GenePT [33], Cell2Sentence [34]	LLM-based approaches for cell-type annotation and natural language-conditioned cell generation.
	Geneformer [36], scGPT [37]	Single-cell foundation models for learning gene/cell representations and batch-effect correction.
	Explainable Sequence Models [45,46]	Integrating sequence with epigenomic maps (ATAC-seq, 3D contacts) to identify brain-specific regulatory elements.
	L2G, PoPS, FLAMES [47,48]	Integrative frameworks combining fine-mapping and multi-omics to prioritize disease candidate genes.
	LINGER [49], GEARS [50], SIGNET [51]	Inference of GRNs and prediction of multigene perturbation outcomes.
	DeepBipolar [52]	Linking mutations to bipolar disorder phenotypes for hypothesis generation.
Boosting psychiatric treatment with AI	SNP-plus-MRI models [53]	Predicting early antidepressant response with high accuracy by combining genomics and neuroimaging.
	Sequence-aware embedding models	Generating variant/gene embeddings to improve patient stratification and drug selection.
	Combinatorial Pharmacogenomics	Outperforming single-gene tests for SSRI exposure prediction using multimodal data integration.

Abbreviations: EHR, electronic health record; FLAMES, fine-mapped locus assessment model of effector genes; GEARS, graph-enhanced gene activation and repression simulator; L2G, locus-to-gene; LINGER, lifelong neural network for gene regulation; PoPS, polygenic priority score; SIGNET, statistical inference on gene regulatory networks; SSRI, selective serotonin reuptake inhibitor.

Autism spectrum disorder (ASD) provides a clear example of this pipeline. Zhou and Troyanskaya [17] first demonstrated that DL-derived variant-effect scores can prioritize noncoding mutations by their predicted regulatory disruption. Building on this, Zhou et al. [55] applied a whole-genome framework to 1790 ASD simplex families and showed that probands harbor *de novo* noncoding mutations with higher predicted functional impact than siblings, converging on neuronal genes and neurodevelopmental pathways. Reporter assays further confirmed allele-specific regulatory activity, supporting causal relevance. These findings illustrate how sequence-to-function predictors (e.g. DeepSEA, Enformer, and ChromBPNet) can rank variants within ASD loci and nominate candidate mechanisms for validation.

Beyond sequence-to-expression prediction, related models extend to epigenetic regulation. For example, DNA methylation (DNAm), a process essential for brain development and implicated in disorders such as schizophrenia [56,57], can be predicted using a transformer-based model that processes 2 kb DNA windows to estimate DNAm levels at CpG sites [41]. This approach enables fine-mapping of regulatory variants affecting DNAm in specific brain cell types, further advancing mechanistic insight into psychiatric disorders.

Multimodal strategies further refine risk estimates by adding transcriptomic or proteomic profiles to genomic features, including work on depression chronicity [58,42]. Emani et al. [7] links genotype to phenotype by integrating scQTL and bulk eQTL sites, cell type-specific and bulk gene regulatory networks (GRNs), cell fractions, cell–cell communication networks, coexpression modules, and sample covariates, thereby anchoring risk to regulatory circuitry and tissue composition. Additional frameworks incorporate electronic health records, clinical variables, and wearable-derived behavior to capture environment and disease course [43,44]. Taken together, DL that respects regulatory biology, coupled with multi-omics and clinical context, promises more accurate and clinically interpretable risk estimates.

Understanding disease mechanisms via artificial intelligence

Understanding disease mechanisms proceeds by anchoring genetic risk in a high-resolution brain cell atlas, mapping noncoding variants to cell type-specific regulatory elements using sequence, single-cell, and 3D genome context with experimental validation, and linking elements to target genes and programs through integrative prioritization and network inference, yielding pathways that shape neural circuits and highlighting therapeutic targets.

Building a cell atlas for psychiatric diseases

A cell-resolved brain atlas provides the cell-, region-, and stage-specific context in which genetic risk manifests in psychiatric disease [59]. As single-cell resources expand (e.g. PsychENCODE) [6,7], consistent annotation and cross-study harmonization become essential; single-cell 'foundation' models trained on large atlases now support automated labeling and integration.

Early strategies repurposed language models for annotation: GenePT prompts ChatGPT with curated gene descriptions to build gene embeddings aggregated into cell profiles [33], while Cell2Sentence reformulates top-expressed gene lists as sentence-like inputs and, after instruction tuning, enables cell-type annotation and natural-language-conditioned cell generation [34]. Purpose-built models (Geneformer, scGPT) learn gene/cell representations from large compendia, improving annotation and mitigating batch effects [36,37]. For psychiatry, atlases emphasizing neuronal and glial lineages, regional specificity, developmental trajectories, and disease-altered states, with cross-cohort validation and interpretable outputs, form the scaffold for mechanism-focused analyses.

Discovering disease-causing molecular variations

Most psychiatric risk variants are noncoding and act through cis-regulatory elements that control gene expression in brain cell types, such as enhancers, promoters, silencers, and insulators. AI integrates native sequence with epigenomic maps (ATAC-seq, ChIP/CUT&Tag, DNAm) and 3D genome contacts to identify brain-specific regulatory elements and assign target genes; typical workflows curate multi-omic datasets, derive motif/accessibility and enhancer–promoter features, and train validated, explainable models [45,46,60].

Applications increasingly focus on neurodevelopmental and glial contexts, quantifying variant effects on transcription factor binding and chromatin state. In single-cell analyses, scRNA-seq and scATAC-seq reconstruct regulatory programs and link distal elements to expression, while 3D contacts refine enhancer-gene maps. *In silico* predictions are paired with clustered regularly interspaced short palindromic repeatsi/a and massively parallel reporter assays to validate effects and prioritize variants with plausible neurobiological mechanisms, enabling pathway- and cell state-level interpretation [45,46,61].

Deciphering regulatory genes in psychiatric diseases

Building on variant-to-regulatory maps, the goal is to connect GWAS signals to genes and transcriptional programs in relevant brain cell types. Summary-data-based Mendelian Randomization and transcriptome-wide Mendelian Randomization link expression quantitative trait loci to candidate genes but are limited by linkage disequilibrium and tissue/cell resolution [62,63]. Integrative frameworks (locus-to-gene, polygenic priority score, and fine-mapped locus assessment model

of effector genes) combine fine-mapping, expression quantitative trait loci, protein quantitative trait loci, and chromatin features to prioritize genes implicated in neurodevelopment, synaptic biology, and glial function [47,6,48,64].

Beyond gene lists, GRN and perturbation-focused approaches clarify the mechanism. Lifelong neural network for gene regulation infers transcription factor activity from bulk and single-cell multi-omics [49]; graph-enhanced gene activation and repression simulator predicts outcomes of multigene perturbations and captures non-additive interactions [50]; dynamic GRN inference with optimal transport recovers temporal regulation [65]; and statistical inference on gene regulatory networks enables transcriptome-wide causal inference [51]. Disease-focused models link mutations to bipolar disorder (DeepBipolar) [52] and implicate noncoding mutations in autism [66], connecting association signals to regulatory mechanisms and target genes for hypothesis generation and therapeutic discovery.

Boosting psychiatric treatment with artificial intelligence

Predicting treatment response is central to precision psychiatry. AI models (neural networks, support vector machines, and random forests) learn non-linear relationships between SNPs, transcriptomes, and outcomes, supporting patient stratification, drug selection, and adverse-effect mitigation [53]. Emerging sequence-aware models generate variant and gene embeddings that capture regulatory context and long-range dependencies, improving feature representation and clinical interpretability.

Performance improves when genomics is paired with polygenic scores, combinatorial pharmacogenomics, DNAm, neuroimaging endophenotypes, demographics, and electronic health record data. For early antidepressant response, SNP-plus-MRI models exceed 80% accuracy [53], and combinatorial panels outperform single-gene tests for selective serotonin reuptake inhibitor exposure. Although clinical use of large sequence models is nascent, their ability to encode context-rich sequences and integrate heterogeneous data promises more accurate, actionable predictions to personalize care.

Conclusion

In this review, we first examine the latest DL techniques that have recently emerged in the field of genomics. Given this context, we then highlight their applications in psychiatric genomics, where many research problems have already actively incorporated these methods, while others are still under development. We hope this review provides a thorough overview of recent advances in DL for genomics and serves as a bridge to help researchers in psychiatric genomics more readily understand and apply these frontier methods to guide the development of novel therapeutic strategies for psychiatric disorders.

Data Availability

No data were used for the research described in the article.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jing Zhang reports financial support was provided by National Institutes of Health. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Institutes of Health, USA [R01HG012572, R01DA063316].

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Lee PH, Anttila V, Won H, Feng Y-CA, Rosenthal J, Zhu Z, Tucker-Drob EM, Nivard MG, Grotzinger AD, Posthuma D, *et al.*: **Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders**. *Cell* 2019, **179**:1469-1482.
2. Li M, Santpere G, Imamura Kawasawa Y, Evgrafov OV, Gulden FO, Pochareddy S, Sunkin SM, Li Z, Shin Y, Zhu Y, *et al.*: **Integrative functional genomic analysis of human brain development and neuropsychiatric risks**. *Science* 2018, **362**:eaat7615.
3. Duman RS, Aghajanian GK, Sanacora G, Krystal JH: **Synaptic plasticity and depression: new insights from stress and rapid-acting antidepressants**. *Nat Med* 2016, **22**:238-249.
4. The 1000 Genomes Project Consortium: **A global reference for human genetic variation**. *Nature* 2015, **526**:68-74.
5. Geschwind DH, Flint J: **Genetics and genomics of psychiatric disease**. *Science* 2015, **349**:1489-1494.
6. Akbarian S, Liu C, Knowles JA, Vaccarino FM, Farnham PJ, Crawford GE, Jaffe AE, Pinto D, Dracheva S, Geschwind DH, *et al.*: **The psychencode project**. *Nat Neurosci* 2015, **18**:1707-1712.
7. Emani PS, Liu JJ, Clarke D, Jensen M, Warrell J, Gupta C, Meng R, Lee CY, Xu S, Dursun C, *et al.*: **Single-cell genomics and regulatory networks for 388 human brains**. *Science* 2024, **384**:eadi5199.
8. Hwang A, Skarica M, Xu S, Coudriet J, Lee CY, Lin L, Terwilliger R, Sliby A-N, Wang J, Nguyen T, Li H, Wu M, Dai Y, Duan Z, Srinivasan SS, Zhang X, Lin Y, Cruz D, Deans PJM, Alvarez VE, Benedek D, Che A, Cruz DA, Davis DA, Hoffman E, Kaye A, Labadorf AT, Keane TM, Logue MW, McKee A, Marx B, Miller MW, Noller C, Montalvo-Ortiz J, Pierce M, Scott WK, Schnurr P, DiSano K, Stein T, Ursano R, Wolf EJ, Huber BR, Levey D, Glausier JR, Lewis DA, Gelernter J, Holtzheimer PE, Friedman MJ, Gerstein M, Sestan N, Brennand KJ, Xu K, Zhao H, Krystal JH, Young KA, Williamson DE, Che A, Zhang J, Girgenti MJ: **Single-cell transcriptomic and chromatin dynamics of the human brain in PTSD**. *Nature* 2025, **643**:744-754.

9. Piwecka M, Rajewsky N, Rybak-Wolf A: **Single-cell and spatial transcriptomics: deciphering brain complexity in health and disease**. *Nat Rev Neurol* 2023, **19**:346-362.

10. LeCun Y, Bengio Y, Hinton G: **Deep learning**. *Nature* 2015, **521**:436-444.

11. Murphy KP: **Machine Learning: A Probabilistic Perspective**. MIT Press; 2012.

12. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD: **Backpropagation applied to handwritten zip code recognition**. *Neural Comput* 1989, **1**:541-551.

13. Hochreiter S, Schmidhuber J: **Long short-term memory**. *Neural Comput* 1997, **9**:1735-1780.

14. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I: **Attention is all you need**. *Adv Neural Inf Process Syst* 2017, **30**.

15. Nguyen E, Poli M, Faizi M, Thomas A, Wornow M, Birch-Sykes C, Massaroli S, Patel A, Rabideau C, Bengio Y, *et al*.: **HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution**. *Adv Neural Inf Process Syst* 2023, **36**:43177-43201.

16. A. Gu, T. Dao, Mamba: linear-time sequence modeling with selective state spaces, *First conference on language modeling*, 2023.

17. Zhou J, Troyanskaya OG: **Predicting effects of noncoding variants with deep learning-based sequence model**. *Nat Methods* 2015, **12**:931-934.

18. Pampari A, Shcherbina A, Kvon EZ, Kosicki M, Nair S, Kundu S, Kathiria AS, Risca VI, Kuningas K, Alasoo K, *et al*.: **ChromBPNet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants**. *BioRxiv* 2025, 2024-12.

19. •• Linder J, Srivastava D, Yuan H, Agarwal V, Kelley DR: **Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation**. *Nat Genet* 2025, **57**:949-961.

Borzoi, a sequence-to-expression model that predicts cell- and tissue-specific RNA-seq coverage directly from DNA sequence. From these predictions, the authors derive statistics to score variant effects across multiple regulatory layers (transcription, splicing, and polyadenylation), showing performance that matches or exceeds state-of-the-art task-specific models on QTL benchmarks.

20. Jaganathan K, Ersaro N, Novakovsky G, Wang Y, James T, Schwartzentruber J, Fiziev P, Kassam I, Cao F, Hawe J, *et al*.: **Predicting expression-altering promoter mutations with deep learning**. *Science* 2025, **389**:eads7373.

21. Chen KM, Wong AK, Troyanskaya OG, Zhou J: **A sequence-based global map of regulatory activity for deciphering human genetics**. *Nat Genet* 2022, **54**:940-949.

22. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, Assael Y, Jumper J, Kohli P, Kelley DR: **Effective gene expression prediction from sequence by integrating long-range interactions**. *Nat Methods* 2021, **18**:1196-1203.

23. •• Avsec Ž, Latysheva N, Cheng J, Novati G, Taylor KR, Ward T, Bycroft C, Nicolaisen L, Arvaniti E, Pan J, *et al*.: **Advancing regulatory variant effect prediction with AlphaGenome**. *Nature* 2026, **649**:1206-1218.

AlphaGenome, a long-context (1 Mb) sequence-to-function model that predicts thousands of genomic tracks at single-base-pair resolution across diverse modalities (expression, initiation, accessibility, histone marks, TF binding, chromatin contacts, and splicing). Trained on human and mouse, it matches or surpasses state-of-the-art methods on 24/26 variant-effect benchmarks and mechanistically recapitulates clinically relevant variant effects near TAL1. Tools are provided for genome-wide track prediction and variant scoring from the sequence.

24. 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**:2112-2120.

25. Zhou Z, Ji Y, Li W, Dutta P, Davuluri RV, Liu H: DNABERT-2: efficient foundation model and benchmark for multi-species genomes. In *Proceedings of the Twelfth International Conference on Learning Representations*; 2024.

26. Fishman V, Kuratov Y, Shmelev A, Petrov M, Penzar D, Shepelin D, Chekanov N, Kardymon O, Burtsev M: **GENA-LM: a family of open-source foundational DNA language models for long sequences**. *Nucleic Acids Res* 2025, **53**:gkae1310.

27. •• Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, Lopez Carranza N, Grzywaczewski AH, Oteri F, Dallago C, Trop E, de Almeida BP, Sirelkhatim H, *et al*.: **Nucleotide transformer: building and evaluating robust foundation models for human genomics**. *Nat Methods* 2025, **22**:287-297.

Nucleotide transformer, a family of DNA foundation models (50M–2.5B parameters) pre-trained on 3202 human genomes plus 850 genomes from diverse species. The models learn context-specific sequence representations that enable accurate predictions in low-data settings, support low-cost fine-tuning across multiple genomics tasks, show unsupervised attention to key genomic elements, and improve genetic variant prioritization.

28. Schiff Y, Kao CH, Gokaslan A, Dao T, Gu A, Kuleshov V: Caduceus: bi-directional equivariant long-range DNA sequence modeling. In *Proceedings of the International Conference on Machine Learning, PMLR*; 2024:43632–43648.

29. •• Nguyen E, Poli M, Durrant MG, Kang B, Katrekar D, Li DB, Bartie LJ, Thomas AW, King SH, Brixi G, *et al*.: **Sequence modeling and design from molecular to genome scale with Evo**. *Science* 2024, **386**:eado9336.

Evo establishes long-context, single-nucleotide-resolution foundation models as a unified approach for prediction and design across the central dogma and genomic scales. By coupling such models with DNA synthesis and genome engineering, this work charts a path toward more accurate variant interpretation and rapid, data-driven biological design, accelerating the ability to engineer living systems.

30. Brixi G, Durrant MG, Ku J, Poli M, Brockman G, Chang D, Gonzalez GA, King SH, Li DB, Merchant AT, *et al*.: **Genome modeling and design across all domains of life with Evo 2**. *BioRxiv* 2025, 2025-02.

31. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, *et al*.: **Training language models to follow instructions with human feedback**. *Adv Neural Inf Process Syst* 2022, **35**:27730-27744.

32. Hou W, Ji Z: **Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis**. *Nat Methods* 2024, **21**:1462-1465.

33. Chen Y, Zou J: **GenePT: a simple but effective foundation model for genes and cells built from ChatGPT**. *bioRxiv* 2024, 2023-10.

34. •• Levine D, Rizvi SA, Lévy S, Pallikkavaliyaveetil N, Zhang D, Chen X, Ghadermarzi S, Wu R, Zheng Z, Vrkic I, et al. Cell2sentence: teaching large language models the language of biology. In *Proceedings of the International Conference on Machine Learning, PMLR*; 2024:27299–27325.

Cell2Sentence, a framework that converts gene expression profiles into ‘ cell sentences, ’ enabling direct fine-tuning of general LLMs for single-cell tasks. With C2S, LLMs can generate biologically plausible cells from type prompts, accurately perform complex cell-type annotation, and produce data-driven textual descriptions. It demonstrates a simple, accessible pathway to repurpose off-the-shelf language models for transcriptomics, bridging NLP and single-cell analysis to support both predictive (annotation) and generative (cell synthesis, text) applications without the need for bespoke model architectures.

35. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, *et al*.: **The Human Cell Atlas**. *elife* 2017, **6**:e27041.

36. Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, Mantineo H, Brydon EM, Zeng Z, Liu XS, *et al*.: **Transfer learning enables predictions in network biology**. *Nature* 2023, **618**:616-624.

37. •• Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, Wang B: **scGPT: toward building a foundation model for single-cell multi-omics using generative AI**. *Nat Methods* 2024, **21**:1470-1480.

scGPT, a generative pre-trained transformer trained on over 33 million single cells that learns biologically meaningful representations of genes and cells. With lightweight adaptation, scGPT achieves strong performance across diverse downstream tasks, including cell-type annotation, batch and multiomic integration, perturbation-response prediction, and gene network inference.

38. Bian H, Chen Y, Dong X, Li C, Hao M, Chen S, Hu J, Sun M, Wei L, Zhang X: **scMulan: a multitask generative pre-trained language model for single-cell analysis**. *Res Comput Mol Biol* 2024, **14758**:479-482.

39. Hao M, Gong J, Zeng X, Liu C, Guo Y, Cheng X, Wang T, Ma J, Zhang X, Song L: **Large-scale foundation model on single-cell transcriptomics**. *Nat Methods* 2024, **21**:1481-1491.

40. Zeng Y, Xie J, Shangguan N, Wei Z, Li W, Su Y, Yang S, Zhang C, Zhang J, Fang N, *et al*.: **CellFM: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells**. *Nat Commun* 2025, **16**:4679.

41. •• Zhou J, Weinberger DR, Han S: **Deep learning predicts DNA methylation regulatory variants in specific brain cell types and enhances fine mapping for brain disorders**. *Sci Adv* 2025, **11**:eadn1870.

INTERACT is a transformer-based model that predicts brain cell-type-specific DNAm with high accuracy and prioritizes regulatory variants whose effects reflect cellular context, enriching heritability for brain traits. Incorporating predicted variant and CpG effects improves fine mapping for schizophrenia, depression, and Alzheimer's disease and maps causal genes to relevant cell types. Importance: demonstrates that DNAm-informed DL can resolve cell type-specific variant effects, advancing variant-to-function interpretation for complex brain disorders.

42. Smeland OB, Busch C, Andreassen OA, Manchia M: **Novel multimodal precision medicine approaches and the relevance of developmental trajectories in bipolar disorder**. *Biol Psychiatry* 2025, **98**:343-353.

43. Lewin G, Abakasanga E, Titcombe I, Cosma G, Gangadharan S: **Artificial Intelligence-enabled Predictive Modelling in Psychiatry: Overview of Machine Learning Applications in Mental Health Research**. Cambridge University Press; 2025:1-7.

44. Liu JJ, Borsari B, Li Y, Liu SX, Gao Y, Xin X, Lou S, Jensen M, Garrido-Martin D, Verplaetse TL, *et al*.: **Digital phenotyping from wearables using AI characterizes psychiatric disorders and identifies genetic associations**. *Cell* 2025, **188**:515-529.

45. Yassin W, Loedige KM, Wannan CM, Holton KM, Chevinsky J, Torous J, Hall M-H, Ye RR, Kumar P, Chopra S, *et al*.: **Biomarker discovery using machine learning in the psychosis spectrum**. *Biomark Neuropsychiatry* 2024, **11**:100107.

46. Winchester LM, Harshfield EL, Shi L, Badhwar A, AlKhleifat A, Clarke N, Dehsarvi A, Lengyel I, Lourida I, Madan CR, Marzi SJ, Proitsi P, Rajkumar AP, Rittman T, Silajdžić E, Tamburin S, Ranson JM, Llewellyn DJ: **Artificial intelligence for biomarker discovery in Alzheimer's disease and dementia**. *Alzheimer's Dement* 2023, **19**:5860-5871.

47. Mountjoy E, Schmidt EM, Carmona M, Schwartzentruber J, Peat G, Miranda A, Fumis L, Hayhurst J, Buniello A, Karim MA, Wright D, Hercules A, Papa E, Fauman EB, Barrett JC, Todd JA, Ochoa D, Dunham I, Ghoussaini M: **An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci**. *Nat Genet* 2021, **53**:1527-1533.

48. Weeks EM, Ulirsch JC, Cheng NY, Trippe BL, Fine RS, Miao J, Patwardhan TA, Kanai M, Nasser J, Fulco CP, Tashman KC, Aguet F, Li T, Ordovas-Montanes J, Smillie CS, Biton M, Shalek AK, Ananthakrishnan AN, Xavier RJ, Regev A, Gupta RM, Lage K, Ardlie KG, Hirschhorn JN, Lander ES, Engreitz JM, Finucane HK: **Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases**. *Nat Genet* 2023, **55**:1267-1276.

49. Yuan Q, Duren Z: **Inferring gene regulatory networks from single-cell multiome data using atlas-scale external data**. *Nat Biotechnol* 2025, **43**:247-257.

50. Roohani Y, Huang K, Leskovec J: **Predicting transcriptional outcomes of novel multigene perturbations with gears**. *Nat Biotechnol* 2024, **42**:927-935.

51. Jiang Z, Chen C, Xu Z, Wang X, Zhang M, Zhang D: **SIGNET: transcriptome-wide causal inference for gene regulatory networks**. *Sci Rep* 2023, **13**:19371.

52. Sundaram L, Rana Bhat R, Viswanath V, Li X: **DeepBipolar: identifying genomic mutations for bipolar disorder via deep learning**. *Hum Mutat* 2017, **38**:1217-1224.

53. Okpete UE, Byeon H: **Challenges and prospects in bridging precision medicine and artificial intelligence in genomic psychiatric treatment**. *World J Psychiatry* 2024, **14**:1148-1164.

54. Benegas G, Batra SS, Song YS: **DNA language models are powerful predictors of genome-wide variant effects**. *Proc Natl Acad Sci* 2023, **120**:e2311219120.

55. Zhou J, Park CY, Theesfeld CL, Wong AK, Yuan Y, Scheckel C, Fak JJ, Funk J, Yao K, Tajima Y, *et al*.: **Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk**. *Nat Genet* 2019, **51**:973-980.

56. Perzel Mandell KA, Eagles NJ, Wilton R, Price AJ, Semick SA, Collado-Torres L, Ulrich WS, Tao R, Han S, Szalay AS, *et al*.: **Genome-wide sequencing-based identification of methylation quantitative trait loci and their role in schizophrenia risk**. *Nat Commun* 2021, **12**:5251.

57. Kozlenkov A, Li J, Apontes P, Hurd YL, Byne WM, Koonin EV, Wegner M, Mukamel EA, Dracheva S: **A unique role for DNA (hydroxy) methylation in epigenetic regulation of human inhibitory neurons**. *Sci Adv* 2018, **4**:eaau6190.

58. Habets PC, Thomas RM, Milaneschi Y, Jansen R, Pool R, Peyrot WJ, Penninx BW, Meijer OC, van Wingen GA, Vinkers CH: **Multimodal data integration advances longitudinal prediction of the naturalistic course of depression and reveals a multimodal signature of remission during 2-year follow-up**. *Biol Psychiatry* 2023, **94**:948-958.

59. Hagenauer MH, Schulmann A, Li JZ, Vawter MP, Walsh DM, Thompson RC, Turner CA, Bunney WE, Myers RM, Barchas JD, *et al*.: **Inference of cell type content from human brain transcriptomic datasets illuminates the effects of age, manner of death, dissection, and psychiatric diagnosis**. *PloS One* 2018, **13**:e0200003.

60. Javaid H, Petrescu CC, Schmunk LJ, Monahan JM, O'Reilly P, Garg M, McGirr L, Khasawneh MT, Al Lail M, Ganta D, Stubbs TM, Sun BB, Vitsios D, Martin-Herranz DE: **The impact of artificial intelligence on biomarker discovery**. *Emerg Top Life Sci* 2025, **8**:89-105.

61. Xu S, Liu J, Zhang J: **scACT: accurate cross-modality translation via cycle-consistent training from unpaired single-cell data**. Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24. Association for Computing Machinery; 2024:2722-2731.

62. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, Montgomery GW, Goddard ME, Wray NR, Visscher PM, Yang J: **Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets**. *Nat Genet* 2016, **48**:481-487.

63. Porcu E, Rüeger S, Lepik K, Agbessi M, Ahsan H, Alves I, Andiappan A, Arindrarto W, Awadalla P, Battle A, Beutner F, Bonder JM, Boomsma D, Christiansen M, Claringbould A, Deelen P, Esko T, Favé M, Franke L, Frayling T, Gharib SA, Gibson G, Heijmans BT, Hemani G, Jansen R, Kähönen M, Kalnapenkis A, Kasela S, Kettunen J, Kim Y, Kovacs P, Krohn K, Kronberg-Guzman J, Kukushkina V, Lee B, Lehtimäki T, Loeffler M, Marigorta UM, Mei H, Milani L, Montgomery GW, Müller-Nurasyid M, Nauck M, Consortium B, Kutalik Zthe eQTLGen Consortium: **Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits**. *Nat Commun* 2019, **10**:3300.

64. Schipper M, de Leeuw CA, Maciel BA, Wightman DP, Hubers N, Boomsma DI, O'Donovan MC, Posthuma D: **Prioritizing effector genes at trait-associated loci using multimodal evidence**. *Nature Genetics* 2024, **57**:323-333.

65. Wamsley B, Bicks L, Cheng Y, Kawaguchi R, Quintero D, Margolis M, Grundman J, Liu J, Xiao S, Hawken N, *et al*.: **Molecular cascades and cell type–specific signatures in ASD revealed by single-cell genomics**. *Science* 2024, **384**:eadh2602.

66. Zhou J, Park C, Theesfeld C, Yuan Y, Sawicka K, Darnell R, Scheckel C, Fak J, *et al*.: **Whole-genome deep learning analysis reveals causal role of noncoding mutations in autism**. *Nature Genetics* 2019, **51**:973-980.