Urecol Journal. Part D:
Applied Sciences
Vol. 5 No. 2 (2025) pp. 52-66
eISSN: 2797-2089

Published by:
KONSORSIUM
LPPM
PTMA
KOORDINATOR WILAYAH JATENG & DIY

A Bibliometric Review of Genomic Prediction Algorithms in Plants : Trends, Collaboration Networks, and Emerging Research Topics

Rizki Darmawan [1,2], **Adi Wijaya** [2], **Prajoko** [3]
[1] Department of Computer Science, Dian Nuswantoro University, Semarang, Indonesia
[2] Department of Health Information Management, Universitas Indonesia Maju, Jakarta, Indonesia 12430
[3] Department of Information Technology, Universitas Muhammadiyah Sukabumi 43113

rizkidarmawan@uima.ac.id

https://doi.org/10.53017/ujas.288
Received: 07/08/2025 Revised: 03/09/2025 Accepted: 02/11/2025

Abstract

Genomic prediction (GP) has emerged as a transformative methodology in biological, health and agricultural sciences, leveraging genome-wide molecular markers to predict complex traits with increasing accuracy. This bibliometric study aims to analyse the development, collaboration patterns, and thematic trends in GP research in plants between 2015 and 2025. A total of 205 open-access research articles were retrieved from the Scopus database using the PRISMA approach and analysed with the R-based Biblioshiny software. The results show a marked increase in publication output since 2018, peaking in 2024, with an annual growth rate of 14.9%. International collaboration is also substantial, with 41.46% of publications involving authors from different countries. The most prolific authors, including Crossa J, Montesinos-López A, and Montesinos-López OA, demonstrate strong collaborative synergies. Leading journals such as Frontiers in Plant Science *and* The Plant Genome *indicate that the field remains strongly rooted in plant genomics and breeding applications. Keyword co-occurrence analysis identified three major thematic clusters: plant breeding and machine learning, AI-based genomics, and statistical prediction systems. Overall, the findings suggest that GP research has evolved into a mature and highly collaborative interdisciplinary field, with a clear shift from conventional statistical approaches toward machine learning and deep learning-driven methodologies. This study provides a systematic map of the intellectual landscape and highlights promising directions for future research in data-driven plant breeding.*

Keywords: *Bibliometrics, Genomic prediction, Plants*

Studi Bibliometrik terhadap Algoritma Prediksi Genomik pada Tanaman: Tren, Jaringan Kolaborasi, dan Perkembangan Topik Penelitiannya

Abstrak

Prediksi genomik (GP) telah muncul sebagai metodologi transformatif dalam ilmu biologi, kesehatan, dan pertanian, dengan memanfaatkan penanda molekuler di seluruh genom untuk memprediksi sifat-sifat kompleks dengan tingkat akurasi yang semakin tinggi. Studi bibliometrik ini bertujuan untuk menganalisis perkembangan, pola kolaborasi, dan tren tematik dalam penelitian GP pada tanaman antara tahun 2015 hingga 2025. Sebanyak 205 artikel penelitian akses terbuka diambil dari basis data Scopus menggunakan pendekatan PRISMA dan dianalisis dengan perangkat lunak Biblioshiny berbasis R. Hasil penelitian menunjukkan peningkatan signifikan dalam jumlah publikasi sejak tahun 2018, dengan puncaknya pada tahun 2024, serta tingkat pertumbuhan tahunan sebesar 14,9%. Kolaborasi internasional juga cukup tinggi, dengan 41,46%

publikasi melibatkan penulis dari berbagai negara. Penulis paling produktif, termasuk Crossa J, Montesinos-López A, dan Montesinos-López OA, menunjukkan sinergi kolaborasi yang kuat. Jurnal terkemuka seperti *Frontiers in Plant Science* dan *The Plant Genome* menunjukkan bahwa bidang ini masih sangat berakar pada genomika tanaman dan aplikasi pemuliaan. Analisis kemunculan bersama kata kunci mengidentifikasi tiga klaster tematik utama, yaitu: pemuliaan tanaman dan pembelajaran mesin, genomika berbasis kecerdasan buatan, serta sistem prediksi statistik. Secara keseluruhan, temuan ini menunjukkan bahwa penelitian GP telah berkembang menjadi bidang interdisipliner yang matang dan sangat kolaboratif, dengan pergeseran yang jelas dari pendekatan statistik konvensional menuju metodologi berbasis pembelajaran mesin dan pembelajaran mendalam. Studi ini memberikan peta sistematis lanskap intelektual serta menyoroti arah yang menjanjikan untuk penelitian masa depan dalam pemuliaan tanaman berbasis data.

Kata kunci: Bibliometrik; Prediksi genomik; Tanaman

1. Introduction

Genomic prediction (GP) has emerged as a transformative methodology in the biological and agricultural sciences, leveraging genome-wide molecular markers to accurately predict complex traits[1], [2], [3]. This approach is crucial in accelerating genetic improvement programs because it allows the identification of genetically superior individuals without the need for time-consuming[4]. This has fundamentally transformed the classical methods[5], enabling faster and more efficient selection. The relevance of GP extends across multiple fields, including its fundamental impact on plant and animal breeding[6] , where it is used to improve yields, disease resistance[7], and adaptation to changing environmental conditions[8], [9]. Furthermore, GP is increasingly relevant in human health, contributing to disease risk assessment and the development of personalized medicine.

The literature demonstrates that GP represents a significant paradigm shift from conventional breeding methodologies. The reason behind this transformative impact is GP's ability to accelerate the breeding process, reduce the duration of artificial selection experiments, and fundamentally separate selection from phenotyping. This separation is a key source of increased efficiency[10]. The rapid development of GP can be traced to rapid advances in molecular techniques, such as high-throughput genotyping and sequencing technologies. These advances have resulted in the availability of abundant and cost-effective genomic data, which in turn enables comprehensive genome-wide marker analysis. This has facilitated the development and widespread adoption of GP methodology, leading to accelerated genetic improvement and substantial reductions in breeding costs[11], [12]. Therefore, GP is not simply an incremental technological advancement, but a fundamental shift in how genetic selection is conceptualized and implemented, with profound implications for addressing critical global challenges such as food security and advancing personalized medicine[13].

The inherently high-dimensional nature of genomic data demands the application of sophisticated statistical and machine learning approaches to extract meaningful insights and make accurate predictions. Therefore, GP is inherently interdisciplinary. Core definitions of GP consistently emphasize its reliance on statistical and machine learning methods for analysing genomic data[14], [15].

A search of the literature reveals that terms such as "statistical modelling," "machine learning," "deep learning," and "bioinformatics" are consistently mentioned alongside traditional biological fields such as "genomics" and "breeding." This pattern clearly demonstrates that GP is not confined to a single discipline but naturally blends computational science, statistics, and genetics. The interdisciplinary foundation of GP implies a growing demand for researchers with hybrid skills, capable of navigating both biological complexity and advanced computational methodologies[16], [17] . It also highlights the importance of collaborative research models that bridge traditional scientific research and facilitate the sharing of diverse expertise, high-dimensional data, and computational resources[18] .

Conducting bibliometric reviews is crucial for systematically mapping the intellectual landscape, identifying emerging research trends, and understanding the key drivers shaping the field of genomic prediction. This analytical approach provides quantitative insights into research outputs, collaboration patterns among researchers and institutions, and significant thematic shifts over time. Such a comprehensive understanding is crucial for charting the field's trajectory and identifying promising avenues for future research and innovation.

However, despite the increasing number of publications on genomic prediction, comprehensive bibliometric analyses that systematically map the research landscape, collaboration networks, and thematic evolution in this field are still limited. Understanding multiple aspects is crucial to identify research trends, influential contributors, and emerging topics in genomic prediction. Therefore, this study conducts a bibliometric review of genomic prediction research to provide a structured overview of this field and highlight potential directions for future research.

2. Methodology

The search strategy was conducted using data sourced from Scopus [19], the world's most comprehensive and trusted resource. The search in this study used the PRISMA method on July 7, 2025, as shown in Figure 1. The query string used was TITLE-ABS-KEY (("genomic" OR "omic") AND ("deep learning" OR "machine learning") AND ("prediction" OR "selection") and ("plant" OR "crops" OR "plants")).

This study only included research articles; other types of articles, such as books, conference papers, or review papers, were not included as primary sources. English was the language of choice for this study, and the paper type used was open access in order to guarantee transparency and reproducibility of bibliometric analysis, allowing all analyzed documents to be freely accessed by researcher. this file and download the Microsoft Word, Letter file.

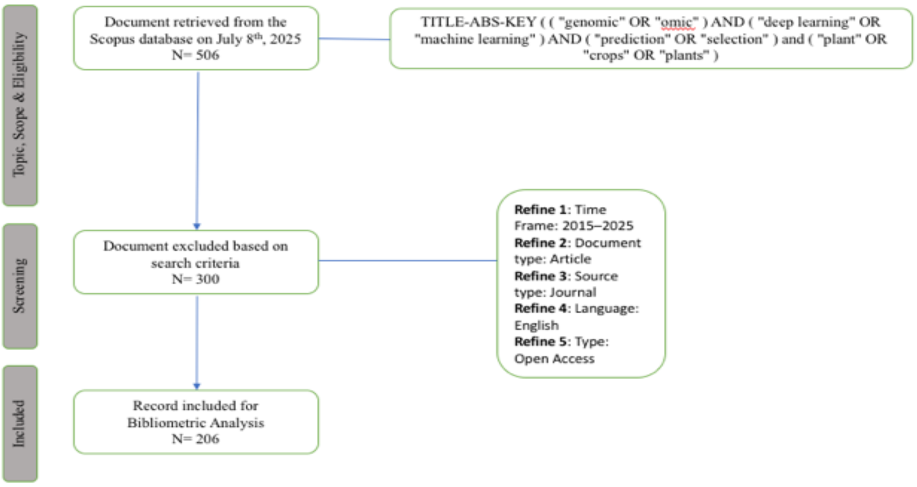

Figure 1. Search strategy based on PRISMA flow diagram

3. Results and Discussions

The annual publication trend for the period 2015 to 2025 is depicted in Figure.2. Based on the data in Figure.2 and Table 1, the number of articles produced each year is relatively low. From 2015 to 2025, the number of articles published was relatively low, often 0. There was a slight spike from 2015 to 2017, with 3-4 articles. A drastic increase was observed starting in 2018. The peak publication rate was in 2024, with 39 articles, marking a significant increase after two years of decline. In July 2025, there were 14 articles. This decrease occurred because data for the current year was incomplete.

Therefore, it can be concluded that the scientific research movement on genomic prediction is limited, as research between 2015 and 2017 indicates minimal or limited research activity. Meanwhile, a significant increase in research began in 2018, peaking in 2024. This increase likely occurred due to several factors, including an increase in the number of researchers, increased research funding, policies encouraging scientific research, new or more productive research focuses, and the maturity of research groups.

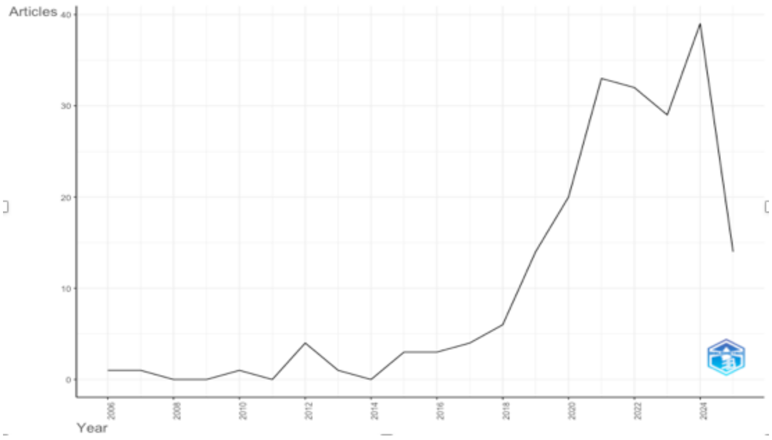

Figure 2. Annual Scientific Production

The fluctuations that occurred from 2021 to 2023 before reaching a new peak in 2024 were not significant. This is within normal limits and could be due to the research project cycle, the length of the journal submission and approval process, or external factors. The drastic decline in 2025 is likely an artifact of incomplete data. It is rare for research that was initially high to decline so drastically the following year unless a major event fundamentally disrupts research activity.

Based on information from Table 2. Main Information, the study period related to genomic prediction research spans the last 20 years, demonstrating continued interest in this topic. There are 205 relevant publications, sufficient for bibliometric analysis. The average annual growth rate of 14.9% indicates significant year-over-year growth in this research. There are 67 different research sources, indicating that research in this field is quite widespread across various scientific publication channels.

The relatively young average age of the documents, 3.69 years, indicates that this topic is highly active and has recently experienced a surge in research. The average 22.26 citations per article indicates a high impact on the scientific community, reflecting the quality and importance of this research. The high number of references, 11,464, indicates that this field is rich in supporting literature and builds on numerous previous studies.

Table 1. Annual publication

Year	Articles
2006	1
2007	1
2008	0
2009	0
2010	1
2011	0
2012	4
2013	1
2014	0
2015	3
2016	3
2017	4
2018	6
2019	14
2020	20
2021	33
2022	32
2023	29
2024	39

Table 2. Main information

Description	Results
Main information about data	
Timespan	2006:2025
Sources (Journals, Books, etc)	67
Documents	205
Annual Growth Rate %	14.9
Document Average Age	3.69
Average citations per doc	22.26
References	11464
Document Contents	
Keywords Plus (ID)	1266
Author's Keywords (DE)	487
Authors	
Authors	1053
Authors of single-authored docs	4
Authors Collaboration	
Single-authored docs	4
Co-Authors per Doc	6.43
International co-authorships %	41.46
Document Types	
Article	205

The large number of keywords per author (487) indicates that this topic is very broad, encompassing approaches, methods, target organisms (e.g., plants), and techniques such as statistics, machine learning, artificial intelligence, and others.

The average of nearly 6.5 authors per document indicates that this research is highly collaborative, reflecting the complexity and interdisciplinary needs of areas such as genetics, data mining, computing, and agronomy.

The high level of international collaboration (41.46%) confirms that this is a global issue with many countries and institutions involved. Research in this field is almost entirely collaborative. Single-authored documents are rare, which is common in experimental and big data-based research fields such as genomic prediction.

The authors with the largest number of authors in this study, as seen in Figure.3 and Table 3, are Crossa J [20], [21], [22], Montesinos-Lopez A, and Montesinos-Lopez OA [23], [24], [25] . All three appear to be members of an active and productive research group, demonstrating close collaboration in genomic prediction research, particularly in the context of agriculture and statistical modeling. Gianola D, despite having only seven publications, recorded the highest average citation count, at nearly 95.3 citations per article, indicating the high quality and influence of each publication. This demonstrates that the number of publications does not always correlate with scientific influence; rather, the quality of the content is also crucial.

All top authors have their most recent publications in 2024, indicating they are still researching and publishing. This is crucial for developing collaborations or searching the current literature. These authors have contributed to numerous collaborative articles, supporting the findings of previous articles that the average citation count is quite high at 6.43 per article and that international collaboration is very dominant in this research topic.

The top researchers mentioned above are core collaborators with a very intense collaboration, as evidenced by over 19 joint publications. The appearance of Hernandes-Suarez[5], [10], [26] as a collaborator supporting the core team indicates a strong collaborative structure, likely based on a multinational institution or project. This can be seen in the data in Table 3.

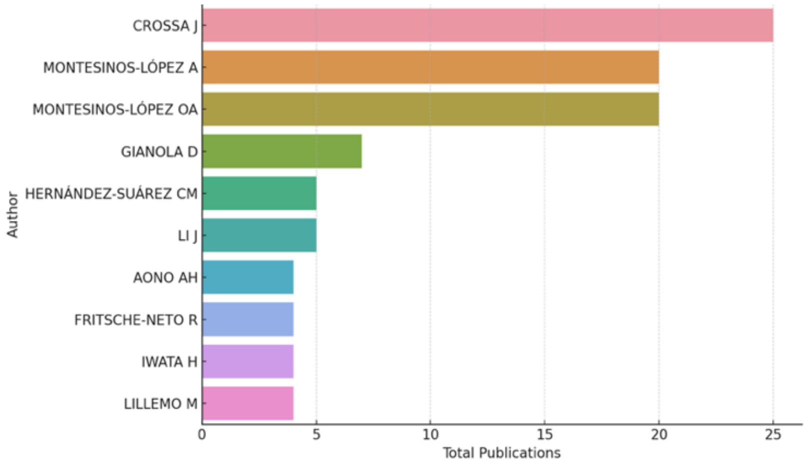

Figure 3. Top 10 Most Prolific Author in Genomic Prediction

Based on Figure.4, it can be seen that the United States dominates with 43 articles, or 20.98% of the total publications. Its relatively low level of international collaboration, at 18.6%, indicates a significant capacity for independent research. Mexico is the country with the highest level of international collaboration, with only four national publications, but 19 international collaborations, representing 82.6% of the total, indicate Mexico's strong connection to global research networks, particularly through authors like Crossa J, who is based at CIMMYT, located in Mexico. Brazil and China also have a strong preference for global research collaboration, with international collaboration scores of 63.3% and 37.9%, respectively.

Table 3. Most productive writers

Author	Publication	Citations Total	Average Citation	Publication Year
Crossa J	25	821	32.84	2024
Montesinos-López A	20	660	33.00	2024
Montesinos-López Oa	20	603	30.15	2024
Gianola D	7	667	95.29	2021
Hernández-Suárez Cm	5	412	82.40	2022

Based on the data from Figure.5, the journal Frontiers in Plant Science is in the top position with 30 articles. This reflects the continued strong focus of predictive genomics research in agronomy and crops. The high thematic relevance found in journal titles like Plant Genome and Frontier in Genetics indicates that publication tends to focus on journals that explicitly address modern genetics. The emergence of journals like Scientific Reports indicates that this topic is also reaching multidisciplinary platforms, beyond just genetics journals.

The dominance of journals such as Frontiers in Plant Science and Plant Genome[27] indicates that GP innovation remains strong in agronomy. However, the emergence of keywords such as "deep learning" and "artificial intelligence" in keyword network analysis indicates a shift from conventional breeding techniques to modern computational approaches. Thematic clusters that have emerge statistical genetics , bioinformatics[28], [29]and machine learning[30], and predictive system confirm this trend.

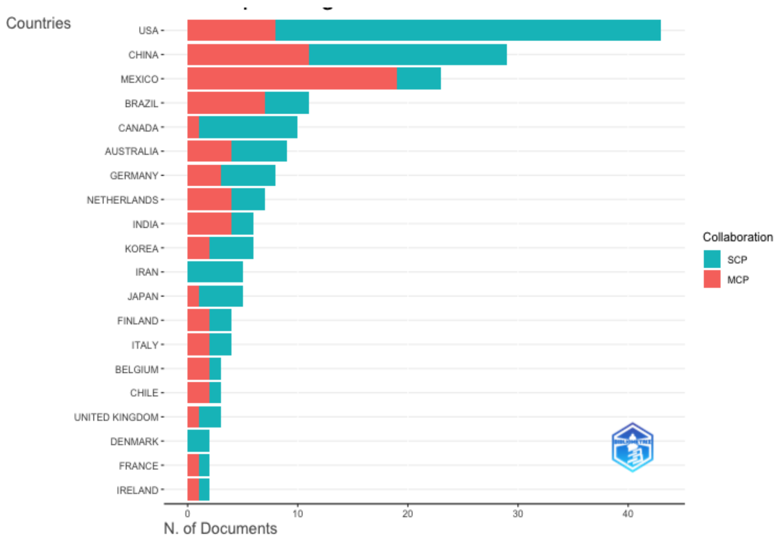

Figure 4. Top 10 countries of origin publications

The high level of international collaboration, particularly from countries like Mexico and Brazil, underscores the global and collaborative nature of GP research. Mexico's dominance of collaboration reflects a reliance on cross-institutional knowledge exchange, particularly through institutions like CIMMYT. This reflects the need for globally shared computing resources and expertise to manage high-dimensional genomic data.

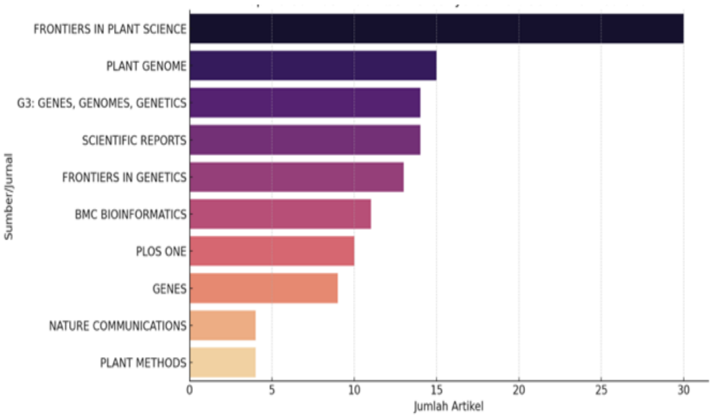

Figure 5. Top 10 journal publications

Table 4 shows that the word "Genomic" is central to the network, indicating that this topic is the foundation of all analyzed research themes. Words such as "deep learning," "genome," and "machine learning" are located near the center, demonstrating the importance of computational approaches in modern genomics research.

Three clusters were identified as seen in Figure.6:

1. Green: Focuses on genetics, plant breeding, phenotyping, and machine learning. This represents classical to modern approaches in plant selection and quantitative trait prediction.
2. Blue: Dominated by genomics, deep learning, artificial intelligence, and genomics. This cluster brings together AI-based approaches and advanced bioinformatics.
3. Red: Related to forecasting, genomic selection, and learning systems. This cluster tends to describe the systematics, prediction, and statistical methodology of genomic prediction.

"Genomic" also has the highest betweenness value of 41.30, indicating that this word acts as a bridge between clusters of ideas. Words with high betweenness often become multidisciplinary topics, connecting different clusters. Current research focuses on the intersection between classical genetics and artificial intelligence technologies.

Table 4. Cluster co-occurance network

Keywords	Cluster	Betweenness	Closeness	PageRank
Genomics	2	41.300	0.020	0.053
Deep learning	2	11.586	0.020	0.025
Genome	2	6.837	0.019	0.020
Bayes theorem	2	1.779	0.018	0.017

Bibliometric studies indicate that genomic prediction (GP) has evolved from a niche research area into a robust, interdisciplinary field integrating genomics [31], machine learning, and statistical modeling. The significant increase in the number of publications since 2018, reaching a peak in 2024, indicates growing global interest and a maturing research ecosystem in this field. Key contributors such as CROSSA J and the MONTESINOS-LÓPEZ group are playing a crucial role in shaping research directions, particularly in plant breeding and agricultural genetics.

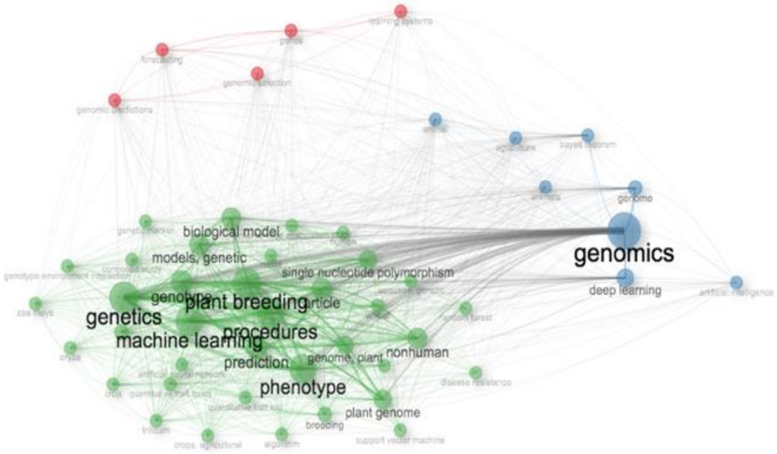

Figure 6. Three cluster co-occurrence network

In the keyword linkage network, the term "genomics" not only appears most frequently but also has the highest betweenness and PageRank values. This confirms that genomics is a key link between research themes. This network structure illustrates the fusion between traditional biological sciences and artificial intelligence technologies a key characteristic of genomic innovation in the future

Figure 7. Word cloud frequent terms

In Figure. 7 and Figure. 8, we can see the dominant themes are machine learning and genomics, indicating that genomic prediction (GP) currently relies heavily on ML approaches. Deep learning [32], [33], [34] is also clearly emerging, indicating a shift from classical statistics to deep learning. The application focus on plant breeding is prominent, especially with the majority of studies oriented towards plant breeding (rather than clinical/human biomedical). The emergence of nonhuman animals, such as Arabidopsis, wheat, and oryza, strengthens the crop and livestock context.

Key data entities in large phenotypes and genotypes demonstrate the core problem of GP, which maps genotypes to phenotypes. The term single nucleotide polymorphism, or SNP [35], confirms commonly used markers. The methods mentioned are specific to support vector machines, random forests, artificial/convolutional neural networks, and Bayes' theorem. The observed shift is Bayes/GBLUP → classical ML → CNN/DL

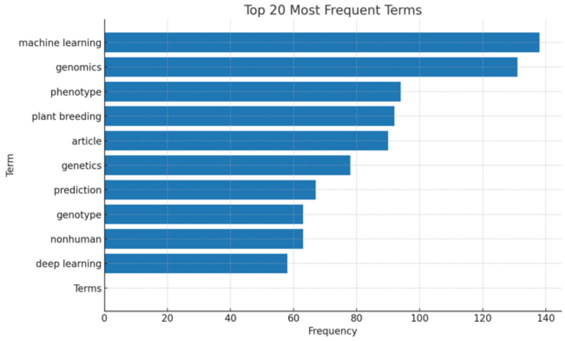

Figure 8. Graphic most frequent terms

In Figure. 9 On the thematic map, the X-axis represents centrality/relevance, and the Y-axis represents density/development. The "machine learning – genetics – genomics" cluster is in the Basic Themes quadrant (bottom right), meaning it is highly relevant to the overall corpus and serves as the foundation for research, but its "density" (internal subtopics) is medium/low—indicating broad themes used across studies.

Several words, such as "deep learning," "nonhuman," and "article," appear in the Niche Themes quadrant (top left): rich in internal connections (dense) but less central to the overall map. This interpretation is Deep learning is actively developing within specific subcommunities, for example, CNNs for certain crops, but not yet as "mainstream" as general ML.

There are almost no large clusters in the Motor Themes quadrant (top right), indicating that there are no subthemes that are both highly central and highly developed; research is likely still scattered between general machine learning (ML) methods and application domains (breeding). The Emerging/Declining quadrant (bottom left) appears to lack meaningful labels on the current plot; to capture new themes, a time-based analysis is required.

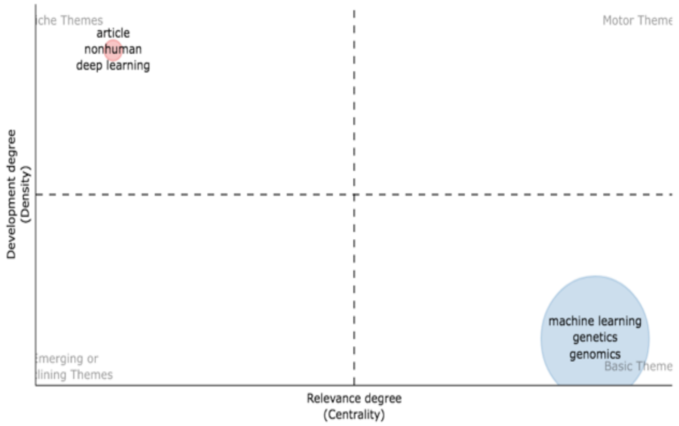

Figure 9. Thematic map.

In Figure. 10 illustrate that the co-citation map shows the core of the Genomic Prediction (GP) literature, highly concentrated around four pillar references: Meuwissen et al. 2001 [36], VanRaden, Endelman and Pérez & de los Campos 2014 [37]. These four papers form hotspots on the density map (the deepest colors), indicating that nearly all articles in the corpus tend to co-cite these works when discussing the basic GP methodology.

The red cluster (Bayesian methodology/Software) centers on Pérez (BGLR package)[38] and the work of de los Campos/Gianola[39]. This is a family of whole-genome Bayesian regression and the software that makes the method easily applicable across studies. Many other nodes are embedded in this cluster, for example, goodfellow Deep Learning which demonstrates the influx of DL, but still as a satellite. The green cluster (BLUP/regularized regression) revolves around VanRaden and embraces statistical learning references such as Friedman, Hastie and Tibshirani and Park & Casella (Bayesian LASSO) [40]. This represents the GBLUP/Ridge/penalized regression pathway, which remains dominant when data size is limited.

The blue cluster (applied genetics/animal–plant breeding) positions Meuwissen as a bridge to applications: along with Goddard/Habier/Crossa and the applied work of 2015–2019. Here, the basic theory of GS flows into plant/livestock breeding studies. The peripheral cluster (purple/orange) contains general ML reference Breiman (RF)[41], Freund & Schapire[42] (Boosting), Kingma & Ba (Adam) [43]as well as some recent work, such as Wang and Clark S.A [44]. Their position is somewhat on the sidelines: useful as a tool, but not central to GP co-citation consensus.

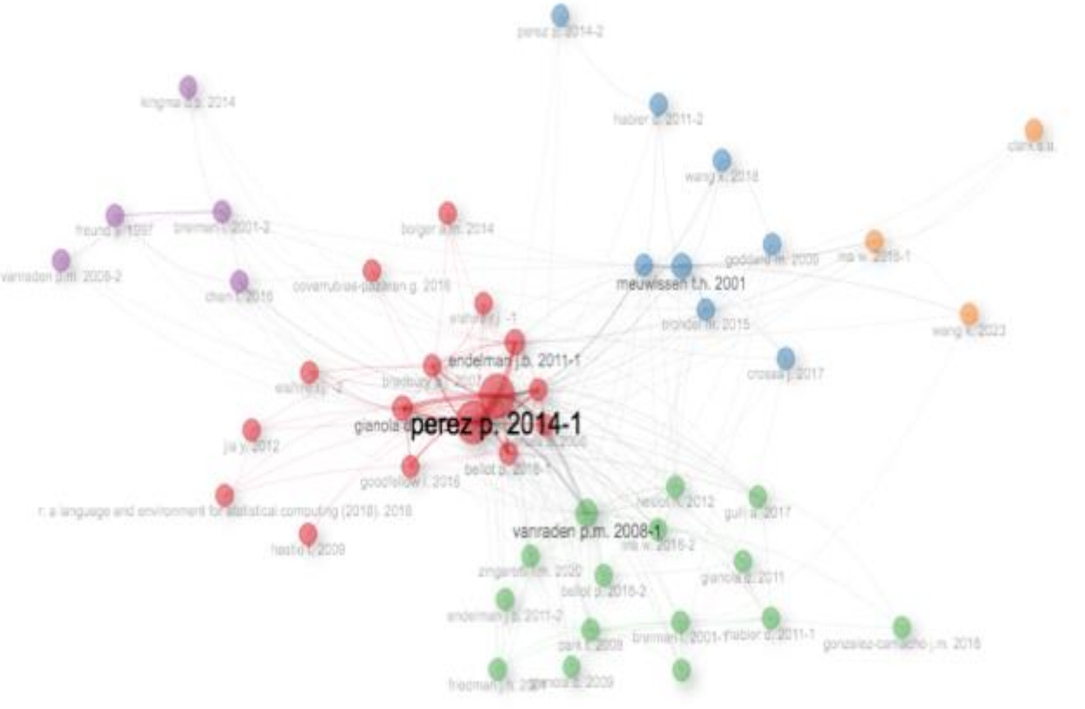

Figure 10. Co-citation map.

4. Conclusion

This study confirms that genomic prediction has become a globally important and collaborative research field, particularly in the crop sciences domain. Between 2015 and 2025, GP demonstrated steady and rapid growth, driven by technological advances and interdisciplinary approaches. The concentration of research output among specific authors and institutions underscores the importance of collaborative networks in knowledge production. The field's intellectual structure is increasingly influenced by data-driven methodologies, where machine learning and artificial intelligence complement conventional breeding practices. The integration of genomics and computational science will be key to addressing global challenges such as food security, climate change, and personalized medicine. Future research should further explore the application of GP to non-crop systems and human health, and strengthen cross-border and cross-disciplinary collaboration through open access publications to foster more inclusive innovation.

Funding

This research received no external funding.

Acknowledgments

The authors would like to express their gratitude to the Department of Health Information Management at Universitas Indonesia Maju, Jakarta and the Department of Information Technology at Universitas Muhammadiyah Sukabumi for providing the academic environment and technical support necessary to conduct this research.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. There are no professional or academic affiliations that may be perceived as a conflict of interest in the conduct of this research or the presentation of its findings.

References

[1] J. Crossa *et al.*, "Genomic prediction of gene bank wheat landraces," *G3: Genes, Genomes, Genetics*, vol. 6, no. 7, pp. 1819–1834, 2016, doi: 10.1534/g3.116.029637.

[2] M. S. Sirsat, P. R. Oblessuc, and R. S. Ramiro, "Genomic Prediction of Wheat Grain Yield Using Machine Learning," *Agriculture (Switzerland)*, vol. 12, no. 9, pp. 1–12, 2022, doi: 10.3390/agriculture12091406.

[3] C. A. V. Barreto *et al.*, "Genomic prediction in multi-environment trials in maize using statistical and machine learning methods," *Sci. Rep.*, vol. 14, no. 1, pp. 1–11, 2024, doi: 10.1038/s41598-024-51792-3.

[4] G. Galli *et al.*, "Automated Machine Learning: A Case Study of Genomic 'Image-Based' Prediction in Maize Hybrids," *Front. Plant Sci.*, vol. 13, 2022, doi: 10.3389/fpls.2022.845524.

[5] A. Montesinos-López, O. A. Montesinos-López, D. Gianola, J. Crossa, and C. M. Hernández-Suárez, "Multi-environment genomic prediction of plant traits using deep learners with dense architecture," *G3: Genes, Genomes, Genetics*, vol. 8, no. 12, pp. 3813–3828, 2018, doi: 10.1534/g3.118.200740.

[6] J. Yan *et al.*, "LightGBM: accelerated genomically designed crop breeding through ensemble learning," *Genome Biol.*, vol. 22, no. 1, pp. 1–24, 2021, doi: 10.1186/s13059-021-02492-y.

[7] C. Palaiokostas, "Predicting for disease resistance in aquaculture species using machine learning models," *Aquac. Rep.*, vol. 20, p. 100660, 2021, doi: https://doi.org/10.1016/j.aqrep.2021.100660.

[8] I. K. Fernandes, C. C. Vieira, K. O. G. Dias, and S. B. Fernandes, "Using machine learning to combine genetic and environmental data for maize grain

yield predictions across multi-environment trials," *Theoretical and Applied Genetics*, vol. 137, no. 8, 2024, doi: 10.1007/s00122-024-04687-w.

[9] J. Cuevas *et al.*, "Deep Kernel for genomic and near infrared predictions in multi-environment breeding trials," *G3: Genes, Genomes, Genetics*, vol. 9, no. 9, pp. 2913–2924, 2019, doi: 10.1534/g3.119.400493.

[10] O. A. Montesinos-López, A. Montesinos-López, J. Crossa, D. Gianola, C. M. Hernández-Suárez, and J. Martín-Vallejo, "Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits," *G3: Genes, Genomes, Genetics*, vol. 8, no. 12, pp. 3829–3840, 2018, doi: 10.1534/g3.118.200728.

[11] P. G. Heilmann, M. Frisch, A. Abbadi, T. Kox, and E. Herzog, "Stacked ensembles on basis of parentage information can predict hybrid performance with an accuracy comparable to marker-based GBLUP," *Front. Plant Sci.*, vol. 14, 2023, doi: 10.3389/fpls.2023.1178902.

[12] A. Montesinos-López *et al.*, "Multimodal deep learning methods enhance genomic prediction of wheat breeding," *G3: Genes, Genomes, Genetics*, vol. 13, no. 5, pp. 1–17, 2023, doi: 10.1093/g3journal/jkad045.

[13] A. Montesinos-López *et al.*, *Deep learning methods improve genomic prediction of wheat breeding*, (2024). doi: 10.3389/fpls.2024.1324090.

[14] O. A. Montesinos-López, A. Montesinos-López, B. Cano-Paez, C. M. Hernández-Suárez, P. C. Santana-Mancilla, and J. Crossa, "A Comparison of Three Machine Learning Methods for Multivariate Genomic Prediction Using the Sparse Kernels Method (SKM) Library," *Genes (Basel).*, vol. 13, no. 8, 2022, doi: 10.3390/genes13081494.

[15] K. Sandhu, S. S. Patil, M. Pumphrey, and A. Carter, "Multitrait machine- and deep-learning models for genomic selection using spectral information in a wheat breeding program," *Plant Genome*, vol. 14, no. 3, 2021, doi: 10.1002/tpg2.20119.

[16] Z. Wang *et al.*, "Optimizing Genomic Selection Methods to Improve Prediction Accuracy of Sugarcane Single-Stalk Weight," *Agronomy*, vol. 14, no. 12, 2024, doi: 10.3390/agronomy14122842.

[17] R. D. S. Rosado *et al.*, "Artificial neural networks in the prediction of genetic merit to flowering traits in bean cultivars," *Agriculture (Switzerland)*, vol. 10, no. 12, pp. 1–12, 2020, doi: 10.3390/agriculture10120638.

[18] K. Wang, M. A. Abid, A. Rasheed, J. Crossa, S. Hearne, and H. Li, "DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants," *Mol. Plant*, vol. 16, no. 1, pp. 279–293, 2023, doi: 10.1016/j.molp.2022.11.004.

[19] R. Pranckutė, "Scopus and Web of Science stands out for systematic reviews, offering comprehensive coverage across disciplines, including journals, conferences, and patents," *Publications*, vol. 9, no. 1, pp. 1–59, 2021.

[20] J. M. González-Camacho, J. Crossa, P. Pérez-Rodríguez, L. Ornella, and D. Gianola, "Genome-enabled prediction using probabilistic neural network classifiers," *BMC Genomics*, vol. 17, no. 1, pp. 1–16, 2016, doi: 10.1186/s12864-016-2553-1.

[21] O. A. Montesinos-López *et al.*, "Multi-Trait, Multi-Environment Genomic Prediction of Durum Wheat With Genomic Best Linear Unbiased Predictor and Deep Learning Methods," *Front. Plant Sci.*, vol. 10, 2019, doi: 10.3389/fpls.2019.01311.

[22] O. A. Montesinos-López, A. Montesinos-López, D. A. Bernal Sandoval, B. A. Mosqueda-Gonzalez, M. A. Valenzo-Jiménez, and J. Crossa, "Multi-trait genome prediction of new environments with partial least squares," *Front. Genet.*, vol. 13, 2022, doi: 10.3389/fgene.2022.966775.

[23] A. Montesinos-López, O. A. Montesinos-López, J. C. Montesinos-López, C. A. Flores-Cortes, R. de la Rosa, and J. Crossa, "A guide for kernel generalized

regression methods for genomic-enabled prediction," *Heredity (Edinb).*, vol. 126, no. 4, pp. 577–596, 2021, doi: 10.1038/s41437-021-00412-1.

[24] O. A. Montesinos-Lopez *et al.*, "Application of a Poisson deep neural network model for the prediction of count data in genome-based prediction," *Plant Genome*, vol. 14, no. 3, pp. 1–19, 2021, doi: 10.1002/tpg2.20118.

[25] O. A. Montesinos-López *et al.*, "A multivariate poisson deep learning model for genomic prediction of count data," *G3: Genes, Genomes, Genetics*, vol. 10, no. 11, pp. 4177–4190, 2020, doi: 10.1534/g3.120.401631.

[26] O. A. Montesinos-López *et al.*, "A benchmarking between deep learning, support vector machine and Bayesian threshold best linear unbiased prediction for predicting ordinal traits in plant breeding," *G3: Genes, Genomes, Genetics*, vol. 9, no. 2, pp. 601–618, 2019, doi: 10.1534/g3.118.200998.

[27] "A foundational large language model for edible plant genomes," *Commun. Biol.*.

[28] H. Kashyap, H. A. Ahmed, N. Hoque, S. Roy, and D. K. Bhattacharyya, "Big data analytics in bioinformatics: architectures, techniques, tools and issues," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 5, no. 1, 2016, doi: 10.1007/s13721-016-0135-4.

[29] S. Ranganathan, "Bioinformatics Education-Perspectives and Challenges," *PLoS Comput. Biol.*, vol. 1, no. 6, pp. 6–8, 2005, doi: 10.1371/journal.pcbi.0010052.

[30] S. Shalev-Shwartz and S. Ben-David, "Understanding machine learning: From theory to algorithms," *Understanding Machine Learning: From Theory to Algorithms*, vol. 9781107057, pp. 1–397, 2013, doi: 10.1017/CBO9781107298019.

[31] R. K. Varshney *et al.*, "Accelerating genetic gains in legumes for the development of prosperous smallholder agriculture: Integrating genomics, phenotyping, systems modelling and agronomy," *J. Exp. Bot.*, vol. 69, no. 13, pp. 3293–3312, 2018, doi: 10.1093/jxb/ery088.

[32] A. Ismail *et al.*, "A Novel Deep Learning-Based Model for Classification of Wheat Gene Expression," *Computer Systems Science and Engineering*, vol. 48, no. 2, pp. 273–285, 2024, doi: 10.32604/csse.2023.038192.

[33] H. Kim, "Deep Learning," *Artificial Intelligence for 6G*, vol. 22, no. 4, pp. 247–303, 2022, doi: 10.1007/978-3-030-95041-5_6.

[34] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[35] F. Jurado-Rui, D. Rousseau, J. A. Botía, and M. J. Aranzana, "GenoDrawing: An Autoencoder Framework for Image Prediction from SNP Markers," *Plant Phenomics*, vol. 5, p. 113, 2023, doi: 10.34133/plantphenomics.0113.

[36] T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard, "Prediction of total genetic value using genome-wide dense marker maps," *Genetics*, vol. 157, no. 4, pp. 1819–1829, 2001, doi: 10.1093/genetics/157.4.1819.

[37] A. Legarra, P. M. VanRaden, E. Mantysaari, M. Bermann, and P. Sullivan, "Scalar methods to deregress and split genomic predictions, and associated behavior of simple regressions, for later use in combined prediction and validations," *J. Dairy Sci.*, vol. 109, no. 2, pp. 1727–1741, 2025, doi: 10.3168/jds.2025-26859.

[38] P. Pérez and G. de los Campos, "Genome-wide regression and prediction with the BGLR statistical package," *Genetics*, vol. 198, no. 2, pp. 483–495, 2014, doi: 10.1534/genetics.114.164442.

[39] D. Gianola, G. De Los Campos, W. G. Hill, E. Manfredi, and R. Fernando, "Additive genetic variability and the Bayesian alphabet," *Genetics*, vol. 183, no. 1, pp. 347–363, 2009, doi: 10.1534/genetics.109.103952.

[40] J. H. Friedman, T. Hastie, and R. Tibshirani, “Regularization Paths for Generalized Linear Models via Coordinate Descent,” *J. Stat. Softw.*, vol. 33, no. 1, pp. 1–22, 2010, doi: 10.18637/jss.v033.i01.

[41] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[42] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997, doi: 10.1006/jcss.1997.1504.

[43] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” 2015. [Online]. Available: https://arxiv.org/abs/1412.6980

[44] S. A. Clark, J. M. Hickey, and J. H. J. van der Werf, “Different models of genetic variation and their effect on genomic evaluation,” *BMC Genet.*, vol. 12, p. 49, 2011, doi: 10.1186/1471-2156-12-49.