



Manifold-based learning for high-throughput single-peanut phenotyping



Weng Kung Peng^{1,2,3} , Xiaomin Lin^{3,4,5}, Peishan Deng^{3,5}, Jianwen Jiang^{3,4}, Chunling Ding^{3,4} & Shijin Yuan³

Peanut (*Arachis hypogaea* L.), a major legume crop valued for its high oil content, displays complex genotypic–phenotypic interactions shaped by environmental influences, yet these relationships remain poorly understood. We present a high-throughput phenotyping framework that captures the geometry of peanut pods using digital microscopy or smartphone imaging integrated with manifold-learning for large-scale analysis and visualization. Using over 6500 pods collected across China, we identify a geographically distinct morphological signature and demonstrate accurate cultivar discrimination. This scalable approach establishes the foundation for a Large Geometric Model capable of predicting phenotypic traits and accelerating precision agriculture. Our pipeline offers a transformative tool for peanut breeding and sustainable crop improvement.

China is the world's leading producer of cultivated peanut (*Arachis hypogaea* L.), yielded 18.3 million tons in 2021, accounting for approximately one-third of global production^{1,2}. As a vital grain legume and oil crop with high protein and oil content, peanut plays a critical role in addressing global food security and poverty. Originating in South America around 6,000 years ago, *A. hypogaea* underwent significant domestication, diverging markedly from its wild ancestors through artificial selection.

Peanuts are classified into two subspecies and six botanical varieties. Subspecies *A. h. fastigiata* exhibits upright growth and shorter crop cycles, while *A. h. hypogaea* displays prostrate growth and longer cycles. Introduced to China by Portuguese traders in the 17th century and later by American missionaries in the 19th century, peanut cultivation in China shows limited genetic diversity, with over 70% of cultivars tracing back to common ancestors. Southern cultivars are predominantly *A. h. fastigiata*, while northern cultivars align with *A. h. hypogaea*^{3,4} (Fig. 1A, B).

A central objective in systems biology is to decipher the complexity of genotypic traits to enhance our understanding of observable phenotypes, aiding plant breeders in crop improvement^{5–12} (Supplementary Table 1). Despite advances in sequencing technologies, predicting phenotypes remains challenging. Overcoming the phenotyping bottleneck in plants (and in general humans, refer to Table 1), hinges on mapping genotypic–phenotypic (GP) correlograms^{5,13}. This requires leveraging high-throughput phenotyping innovations, such as multi-scale, multi-omics, and multi-dimensional approaches, alongside advanced bioinformatics techniques, including manifold learning and dimensional reduction. These tools enable visualization and comprehension of the meta-state space, facilitating a deeper understanding of genotypic–phenotypic relationships and driving precision breeding strategies.

In this work, we demonstrated the application of ubiquitous technologies (e.g., digital microscopy) and advanced data analytics (e.g., manifold learning) to characterize the internal and external phenotypes of individual peanut pods. We analyzed the complex geometries of 6,500 peanut pods across China and we identified for the first time a geographically unique ‘molecular thumbprint’ derived from the topological geometry of single peanut pods. Additionally, we established correlations between biophysical and biochemical traits through multi-factorial phenotyping. These innovative techniques in complementary to traditional approaches such as molecular pathway analysis and gene sequencing, offer significant potential for elucidating molecular/cellular states, and advancing our understanding of complex traits interactions in peanuts^{14–17}.

Results

Overview of the study

This study examines the complex traits of peanut pods (intricate geometry, biophysical and biochemical traits), due to complex interactions among other internal factors (e.g., lipids, proteins) and biotic/abiotic environmental variables. A total of 6,500 peanut pod samples were procured through the Taobao online platform, originating from diverse geographical regions across China, including Yunnan (YN), Guangdong (GD), Hunan (HN), Shandong (SD), Dongbei (DB), and nearby regions (see Fig. 1A and Supplementary Fig. 1 for details). To maintain confidentiality, samples were labeled by region (e.g., YN, GD) without revealing commercial brand names. Based on their geographical origins, the samples were grouped into three regions: South (YN, GD), North (HN, SD), and Northeast (DB), as depicted in Fig. 1B. This classification enables the analysis of regional environmental impacts on peanut pod traits.

¹Institute for Advanced Co-Creation Studies, The University of Osaka, Osaka, Japan. ²Center for Quantum Information and Quantum Biology, The University of Osaka, Osaka, Japan. ³Songshan Lake Materials Laboratory, Dongguan, China. ⁴Guangdong Medical University, Dongguan, China. ⁵These authors contributed equally: Xiaomin Lin, Peishan Deng. ✉e-mail: peng.weng.kung.qiqb@osaka-u.ac.jp

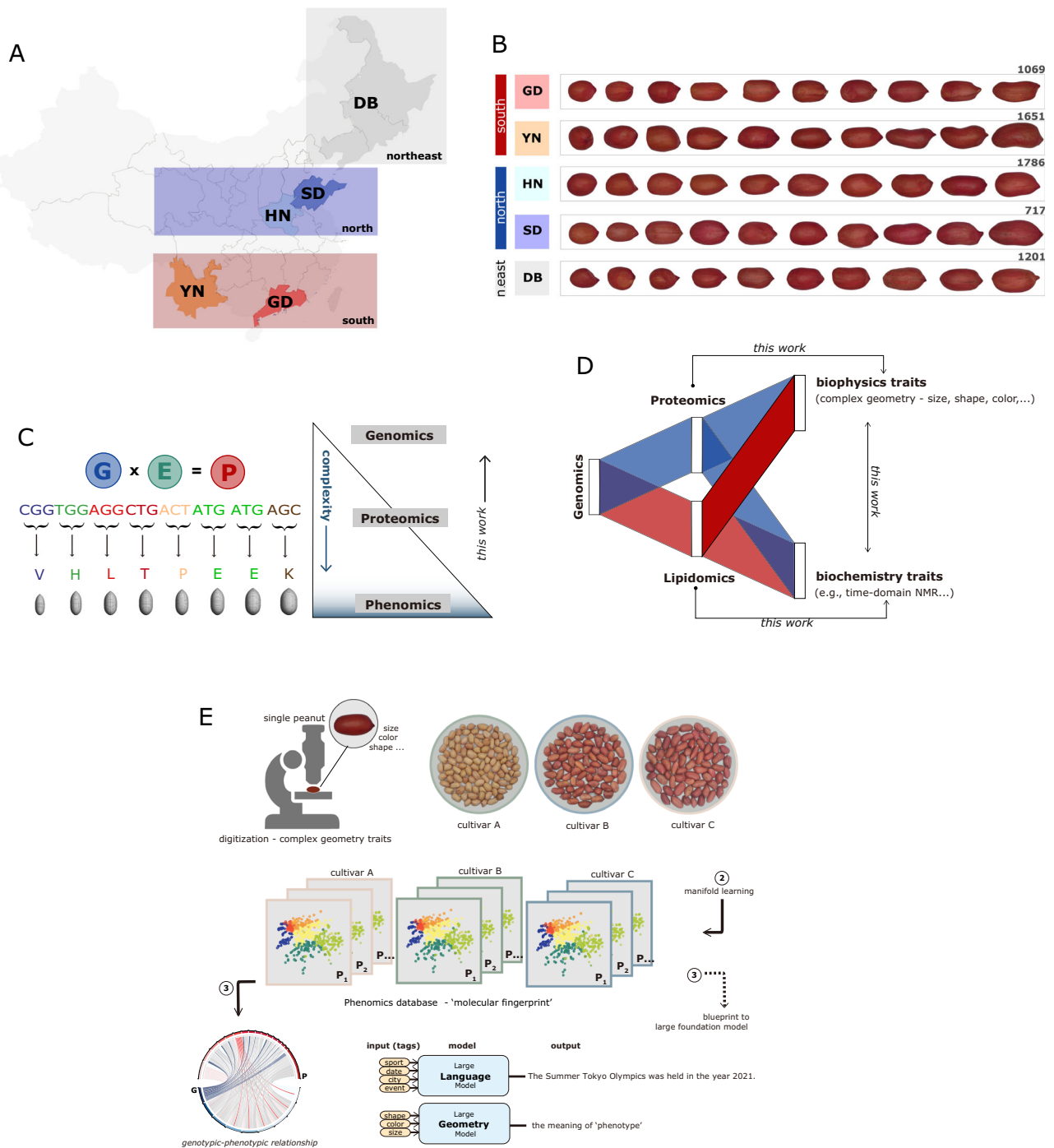


Fig. 1 | Manifold-based learning for high-throughput phenotyping of single peanut pods. **A** Peanut pods cultivated in distinct regions of China: southern (Yunnan, YN; Guangdong, GD), northern (Henan, HN; Shandong, SD), and northeast (Dongbei, DB). **B** Digitized micrographs of off-the-shelf peanut pods from China, with the number of pods analyzed indicated in the top right of each box (details in Supplementary Fig. 1). **C** The complex genotypic-phenotypic (GP) network, modulated by environmental factors E, contributes to the biodiversity of peanuts. Multi-level biological scales (e.g., genomics, phenomics) amplify the dimensional complexity in life sciences. Key environmental factors, including location (latitude/longitude), rainfall, and temperature, directly influence the dynamic GP relationship (Supplementary Information). **D** Correlation mapping between biophysical traits (e.g., length, weight, color) and biochemical traits (e.g.,

time-domain micro-NMR, optical microscopic analysis) of peanut pods from various Chinese regions. **E** Deep and high-throughput characterization of peanut pod phenotypes as proposed in this work. The process of digitization was carefully carried out on each single peanut to obtain its complex geometry (e.g., shape, size, color) and then subsequently fed into the machine learning platforms. These peanuts purchased off-the-shelf via on-line platform were used without any further processing. The intricate Genotypic-Phenotypic relationships were mapped out with Circos. The proposed method has the potential to serve as a blueprint for a large foundation model for the prediction of phenotype based on the complex geometry traits (known as Large Geometric Model), analogous to Large Language Model (LLM) used to generate semantically and grammatically correct sentences.

Table 1 | Comparison of Traditional Genomics-based Medicine and Emerging Phenomics-based Medicine

	Current Genomics-based medicine	Future Phenomics-based medicine
<i>Nature</i>	forward (or reverse) genetics research	big data - phenotypes (e.g., clinical data)
<i>Methodology</i>	genomics-based approach (e.g., genes sequencing, gene editing)	precise measurement of phenotypes (e.g., mass spectrometry, NMR, ...)
<i>Samples</i>	genetic materials (e.g., DNA, SNP, ...)	biofluids (e.g., sweat, plasma), imaging, clinical symptoms
<i>Stability</i>	inherently stable (but static information)	dynamics (fluctuations depending on interactions)
<i>Associations</i>	GWAS-based associations (hypothesis-driven)	PheWAS-based associations (data-driven)
<i>Advantageous</i>	clear relationship between a mutation and disease phenotype	cross-modalities reduces bias, much higher sensitivity (repeated profiling), further sub-stratification possible
<i>Disadvantageous</i>	direct biological, causal complicated inferences, direct linkage impossible	high-cost (large biodiversity, multi-omics), inherent variability among individuals
<i>Information</i>	massive ancestry information (difficulty in interpretation)	G x E (environmental aspects), cross-scale multi levels (deep phenotyping)
<i>Applications</i>	estimating its heritability, calculating genetic correlations, making clinical risk predictions, informing drug development programmes	leading to personalized medicine (e.g., wearable technologies, continuous glucose monitoring)
<i>Clinical use</i>	accurate (esp.monogenic diseases) but limited for chronic disease (due to various lifestyles)	reflective of actual state (further sub-phenotypes possible)
<i>ref(s)</i>	Guttmacher ^a , Perkins ^b , Hoh ^c	Yurkovich ^d , Yan Ling ^e , Hoh ^f , Halama ^g

^aGuttmacher et al, <https://doi.org/10.1056/NEJMra012240>^bPerkin et al, <https://doi.org/10.1073/pnas.1706096114>^cHoh et al, <https://doi.org/10.1038/s41525-024-00435-7>^dYurkovich et al, <https://doi.org/10.1038/s41576-023-00674-x>^eYan Ling et al., <https://doi.org/10.1186/s12864-019-6363-0>^fHoh et al., <https://doi.org/10.15302/J-QB-021-0251>^gHalama et al., <https://doi.org/10.1038/s41467-024-51134-x>

This SWOT-like analysis based on the methodologies, technologies, and outcomes of genomics-based approaches with those of phenomics-based approaches in medical research and practice.

The study employed deep phenotyping to analyze the biochemical and biophysical traits of individual peanut pods, utilizing technological innovations such as light microscopy and micro-NMR^{18–20}, alongside sophisticated bioinformatics tools like manifold-learning. These methods enabled detailed assessment of both internal and external phenotypes of single peanut pods (Fig. 1C–E). Additionally, correlations were investigated between biophysical traits (e.g., length-weight, length-color, weight-color) and biochemical traits (e.g., NMR-based²¹) of pods sourced from various regions across China, with results presented in Table 1, and Supplementary Table 2.

Biophysical traits measurements

Biophysical traits of individual peanut pods, including length, weight, and color, were measured with precision electronics. Length was determined using a Digimatic Caliper (SanLiang Corporation, China), and weight was measured with a high-precision electronic scale (LeaPing, China). Pod color, obtained from digital microscopy, was compared to a standardized color database (see Methods and Supplementary Fig. 2 for details).

Significant phenotypic variation in biophysical traits (length, weight, color) was observed across the 6,500 peanut pod samples (Fig. 2A). Strong positive correlations between pod weight and length were identified across regions: Guangdong (GD: $r = 0.86$, $R^2 = 0.73$), Yunnan (YN: $r = 0.66$, $R^2 = 0.43$), Hunan (HN: $r = 0.67$, $R^2 = 0.45$), Shandong (SD: $r = 0.61$, $R^2 = 0.38$), and Dongbei (DB: $r = 0.64$, $R^2 = 0.41$), where r represents the Pearson correlation coefficient and R^2 the coefficient of determination (Fig. 2B). Classification efficiency of biophysical traits was evaluated using receiver operating characteristic (ROC) analysis (Supplementary Table 2).

Notably, pods from the same region (e.g., South: YN, GD; North: HN, SD) exhibited greater similarity in biophysical traits compared to those from different regions (Fig. 2C). Interestingly, both length and weight traits exhibit region-specific patterns (e.g., North vs. South), as evidenced by statistically significant effects in the analysis of variance (ANOVA) test ($p < 0.001$, see Methods). Color analysis treated each color pod as a single image in a Red-Green-Blue (RGB) ternary plot, with relationships sorted using t-distributed stochastic neighbour embedding (t-SNE) analysis. At

lower perplexity (e.g., <5), pod color closely aligned with geographical origin, showing localized distribution in the ternary plot, whereas higher perplexity (e.g., >10) resulted in less distinct clustering.

Complex Geometry and Manifold Learning – t-SNE Analysis

The interplay between genetic factors (G) and diverse environmental conditions (E) generates significant biodiversity in phenotypic traits²². This study hypothesizes that the complex geometry of peanut pods creates a unique ‘molecular thumbprint’ specific to their cultivation region, serving as a signature of their origin^{6,9} (Fig. 2 and Fig. 3).

To capture the biophysical traits of each peanut pod, a low-cost digital microscope (Shengchuang 1080 FHD, China) with image resolution (2400×1800 pixels) and the lighting condition controlled inside a photography box (AMBITFUL, China) (see Methods, with calibration details in Supplementary Fig. 3 and Supplementary Fig. 4). Beyond the standard three-dimensional (3D) physical structure, individual biophysical traits (e.g., size, shape, color) were treated as distinct dimensions (Fig. 2). Digitization enabled manipulation of these dimensions, facilitating access to higher-dimensional (meta-dimensional) data and subsequent reducing the observable dimensions using advanced techniques^{23,24}. Manifold learning, specifically t-SNE analysis, was applied to explore the complex geometry of these traits. This approach revealed patterns in the high-dimensional data, highlighting how phenotypic variations, such as color and shape, correlate with geographical origins, reinforcing the hypothesis of region-specific molecular thumbprints.

Using t-SNE analysis, the peanut pod samples were predominantly clustered into two distinct subgroups based on their biophysical traits: the southern region (red, right) and the northern region (blue, left), as shown in Fig. 3A, where each point represents an individual peanut pod. Further analysis revealed finer subgroupings within these regions, aligning with geographical proximity: the southern subgroup included Yunnan (YN) and Guangdong (GD), while the northern subgroup comprised Hunan (HN) and Shandong (SD) (Fig. 3B and Supplementary Fig. 5A). The northeast region (Dongbei, DB, grey) exhibited a more uniform distribution, though it showed a slight skew toward the southern region, indicating some phenotypic overlap.

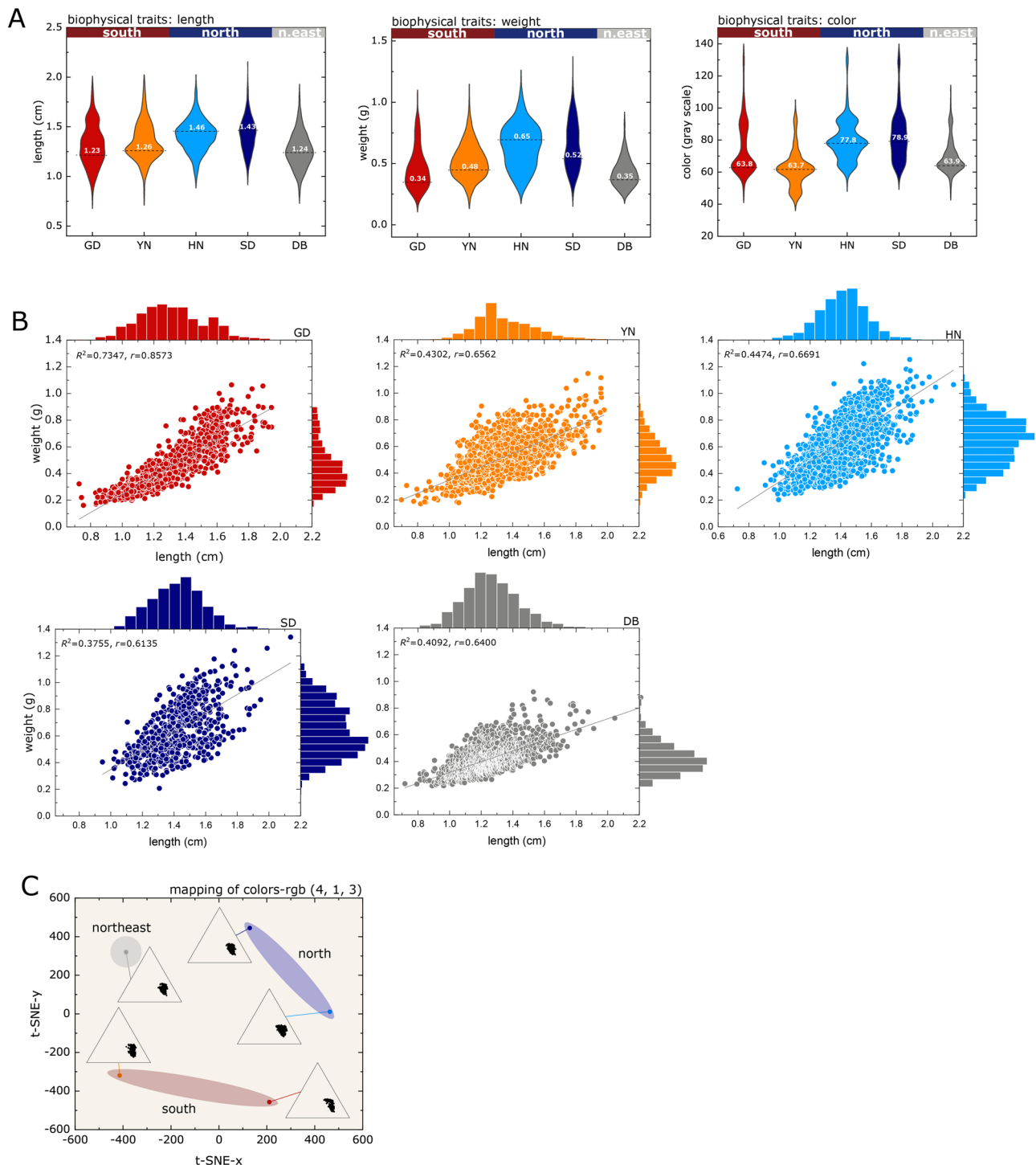


Fig. 2 | Correlational mapping of biophysical traits of peanut pods. **A** Distribution of measured biophysical traits (length, weight, greyscale color) of peanut pods from various cultivar regions in China: Guangdong (GD), Yunnan (YN), Henan (HN), Shandong (SD), and Dongbei (DB). **B** Correlation analysis between length and weight of peanut pods, with Pearson correlation coefficient (r) and determination

coefficient (R^2) indicated on the graph. Analysis of variance (MANOVA) revealed statistically significant effects for all tested cases ($p < 0.001$) (details in Methods). **C** Correlational analysis of RGB colour traits of peanut pods using t-SNE analysis (details in Supplementary Fig. 3).

At lower perplexity settings (e.g., t-SNE (10, 1, 2)), t-SNE analysis revealed distinct ridges and ponds in the distribution of peanut pod biophysical traits, forming continuous trails that segmented the data (Fig. 3A and Fig. 3C). These patterns are hypothesized to result from the presence or absence of specific pod geometries tied to cultivation regions. The predominant geometries observed were combinations of round and rectangular shapes, with intermediate forms, while perfectly round or perfectly

rectangular pods were rare. These ridges, unique to each region, form a 'molecular thumbprint' reflecting the composite influence of environmental modulators on pod geometry, varying by geographical location (Fig. 3B). Similar patterns were observed at t-SNE (30, 1, 2).

Phenotypic trends showed smaller, darker peanut pods clustering predominantly on the left side of the t-SNE plot, with pod size increasing toward the right (Fig. 4). These traits underscore the biodiversity arising

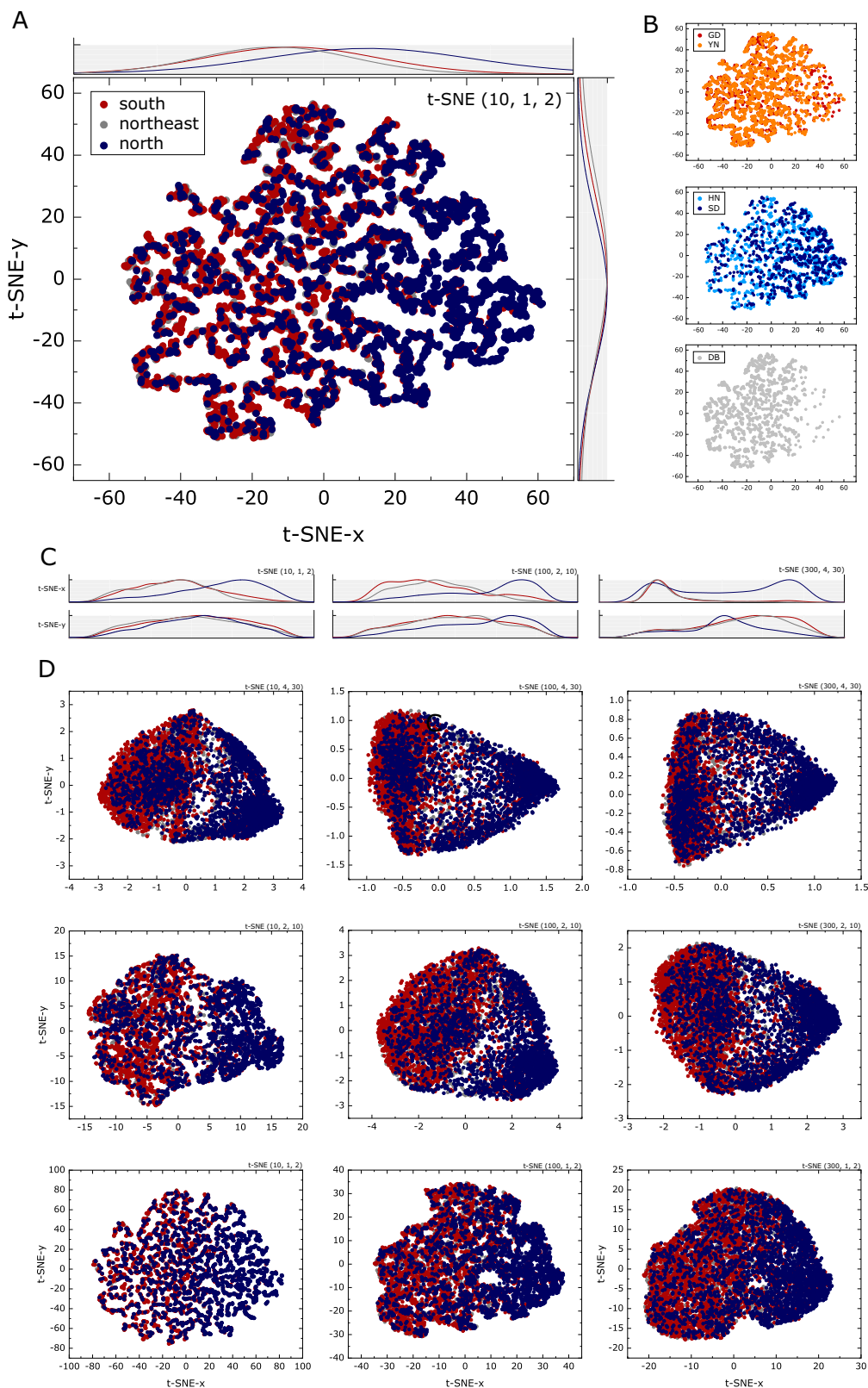


Fig. 3 | Digitized molecular thumbprint of single peanut pods using t-SNE analysis. A Phenotypic mapping of peanut pods from South (red) and North (blue) regions of China, constructed from digitized images of complex pod geometries processed via manifold-based learning with t-SNE (Perplexity=10, Exaggeration=1, PCA = 2) (details in Supplementary Fig. 1). Each dot represents a single peanut pod image. B Detailed mapping by cultivar regions: Guangdong (red), Yunnan (orange),

Henan (blue), Shandong (dark blue), and Dongbei (grey), using the same t-SNE parameters (10, 1, 2) as in A. C Cross-sectional distribution along the X-axis (top) and Y-axis (bottom) for t-SNE configurations (10, 1, 2), (100, 2, 10), and (300, 4, 30). D Impact of t-SNE parametric variations on phenotypic mapping, with parameters ranging across Perplexity (1–500), Exaggeration (1–4), and PCA (2–50).

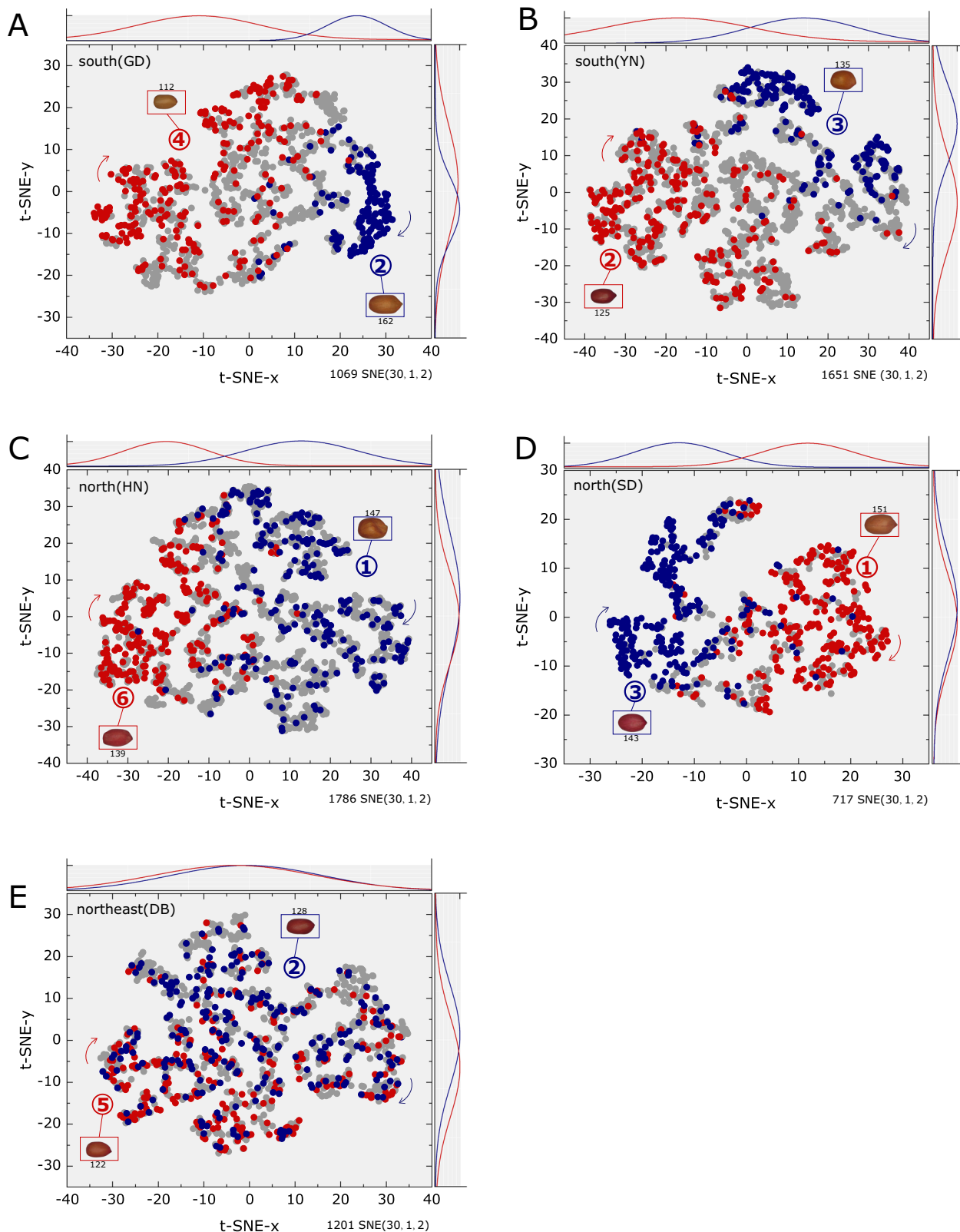


Fig. 4 | Geographical-based molecular thumbprint of peanut pods cultivated across China. This figure maps the complex geometry of single peanut pods from five regions: **A** South (Guangdong, GD), **B** South (Yunnan, YN), **C** North (Henan, HN), **D** North (Shandong, SD), and **E** Northeast (Dongbei, DB). Each point represents an individual peanut pod image, with the total number of pods indicated on the right side of each frame. Points are color-coded by sub-locations (blue and

red) and other locations (grey), with numbering randomly assigned based on sub-locations (details in Supplementary Fig. 1). t-SNE analysis was applied to reduce dimensionality into t-SNE-x and t-SNE-y axes, using parameters (Perplexity, Exaggeration, PCA). The number of peanut pods analyzed is denoted below each frame.

from complex gene-environment interactions, described as ‘deterministic biology’^{16,25}. At higher perplexity (e.g., t-SNE (300, 4, 30)), the local structure collapsed, forming broad peaks representing the southern and northern regions. This ‘swarm of bees’ pattern reflects collective behaviour rather than localized structures, as shown in Fig. 3D and Supplementary Fig. 5B, highlighting the interplay between regional environmental factors and phenotypic variation.

Simulation: Complex geometry analysis of peanut pods

This study hypothesized that the observable physical traits of peanut pods, such as their complex geometry (e.g., size, shape), influence the demographic patterns in distribution charts, particularly clustering behaviour. Mathematical models were developed to simulate these physical traits, focusing on size and shape, followed by manifold-based learning analysis using t-SNE. The results showed that size-based analysis produced continuous patterns; while shape-based analysis yielded non-continuous patterns (see Methods and Supplementary Fig. 6A). This distinction is attributed to size variations likely representing intraspecies differences, whereas shape variations suggest interspecies differences. Consistent patterns were observed across various t-SNE parameter combinations (Supplementary Fig. 6B–6C), reinforcing the impact of geometric traits on clustering and distribution outcomes.

The complex interplay between genetic factors (G) and diverse environmental variables (E) generates a wide array of phenotypic traits, contributing to biodiversity^{22,25}. Mathematically, the resulting molecular thumbprint, P, can be represented as a high-dimensional array of matrices modulated by environmental factors, E (e.g., E1, E2, ...). In the era of artificial intelligence, techniques such as deep learning and machine learning, coupled with manifold learning (e.g., dimension reduction), enable the deciphering and visualization of these meta-dimensional patterns, facilitating the interpretation of complex phenotypic data.

Manifold Learning - Uniform Manifold Approximation and Projection (UMAP) Analysis

In contrast to t-SNE, the UMAP technique provides superior preservation of global data structure^{23,24}, as illustrated in Fig. 5. UMAP analysis distinctly separated peanut pod samples into subgroups corresponding to their cultivation regions: Guangdong (GD) and Yunnan (YN) clustered together, as did Hunan (HN) and Shandong (SD). This indicates that peanut pods from regions sharing similar latitudes exhibit comparable phenotypic traits (Fig. 5A, B). The GD-YN subgroup represents the southern regions (red, orange), while the HN-SD subgroup corresponds to the northern regions (dark blue, blue). This clustering aligns with expectations, as geographically proximate regions often share similar climatic conditions, temperature, and rainfall, which are critical environmental factors influencing genotypic-phenotypic correlations. Notably, geographical regions are frequently defined by political boundaries rather than scientific or environmental criteria, as detailed in Fig. 6.

UMAP analysis revealed two significant phenomena. First, peanut pods from Guangdong (GD) occupied a distinct area in the phenotypic distribution, not shared by other regions, indicating a unique phenotype³. This is likely due to Guangzhou’s role as a gateway to international trade, fostering greater biological diversity^{2,26}. Second, a small but notable proportion of Hunan (HN) phenotypes appeared within the southern region of the UMAP mapping. This could reflect HN’s prominence as a well-known and popular region in China, making its peanut varieties a focal point for breeders^{2,26}. Overall, the UMAP-based mapping results align closely with those from t-SNE, confirming consistent phenotypic clustering across both manifold learning techniques.

Geographically specific molecular thumbprints

To substantiate the claim of geographically unique molecular thumbprints derived from peanut pod geometries across multiple locations, we evaluated the robustness of clustering patterns observed in manifold-based learning techniques (e.g., t-SNE) to ensure cluster stability and statistical significance

(see Methods). Specifically, k-fold cross-validation ($k = 10$), leave-one-out validation (LOOV), and stratified random sampling were employed to mitigate potential stochasticity in the embeddings. Downstream classification performance was quantified using the area under the receiver operating characteristic curve (AUC) across multiple algorithms, including k-nearest neighbours (kNN), support vector machines (SVM), neural networks, logistic regression, and random forests. The results demonstrate high predictive accuracy, with median AUC values exceeding 0.986 for direct image embeddings, 0.931 for t-SNE-derived features, and 0.900 for Isomap projections, thereby underscoring the geographic specificity of the fingerprints (Supplementary Fig. 10).

Modulation Effect of Environmental Factors on the Molecular Thumbprint

To investigate the relationship between the molecular thumbprint of peanut pods and their geographical origins, phenotypic patterns were analyzed using t-SNE (Fig. 6A). Environmental factors, such as temperature, rainfall, and soil conditions, significantly influence genotypic-phenotypic correlations. As expected, t-SNE analysis formed two distinct clusters based on geographical proximity: Guangdong (GD) and Yunnan (YN) in the southern region, and Hunan (HN) and Shandong (SD) in the northern region. These findings were consistent across various t-SNE parameter combinations (Supplementary Fig. 7A, 7B, 7C) and corroborated by hierarchical clustering analysis (Fig. 6B).

Peanut cultivation in China spans diverse climates, from the frigid northern regions to the humid southern regions, with temperatures in peanut-producing areas ranging from -5°C to 25°C (Fig. 6). However, temperature and rainfall exhibit day-to-day variability and form distinct, though similar, annual patterns. A manifold-based learning approach applied to temperature and rainfall data (Fig. 6C, D, and Supplementary Fig. 7D) similarly separated samples into southern and northern clusters, reflecting shared environmental conditions like temperature, rainfall, and soil properties within proximate regions.

These phenotypic and environmental patterns align with genotypic observations reported by Ren et al.³. All southern samples (GD, YN) belonged to the cultivar *ssp. fastigiata*, while northern samples (HN, SD) were of *ssp. hypogaea*, demonstrating consistent clustering across

Geographical-based Molecular Traits with Micro-NMR Analysis

To explore the correlation between biophysical traits (e.g., length) and biochemical traits (e.g., lipidomics), micro-NMR relaxometry experiments, including T_1 and T_2 relaxation measurements, were conducted. Oil droplets were extracted from individual peanut pods using a cold-pressed technique, with proton nuclei relaxation rates primarily reflecting lipid content (e.g., saturated and unsaturated fats). Each peanut type was sampled five times, with three measurements per sample (see Methods for details).

As anticipated, the greatest discrimination in NMR-based molecular traits was observed between the northern and southern regions of China (highlighted in red), while discrimination was lower (blue) among samples within the same region, as shown in the heatmap (Supplementary Table 2 and Supplementary Fig. 8). The northeastern region, with the exception of LN₁, appeared as a natural extension of the northern regions based on NMR-derived molecular traits, indicating biochemical similarities driven by geographical and environmental factors. Interestingly, the T_2 traits ($p < 0.0001$, Student’s t-test) exhibit geographically specific properties, in contrast to the T_1 traits ($p = 0.1$, Student’s t-test) (Supplementary Information).

Phenotype-genotype mapping network based on single peanut pod typing

This study extends beyond correlations within internal phenotypes (e.g., biochemical vs. biochemical traits) and external phenotypes (e.g., biophysical vs. biophysical traits) to establish relationships between internal and external phenotype spaces (i.e., biochemical vs. biophysical traits), as summarized in Supplementary Table 2. Notably, strong positive correlations were observed between T_1 and T_2 relaxation rates (biochemical traits)

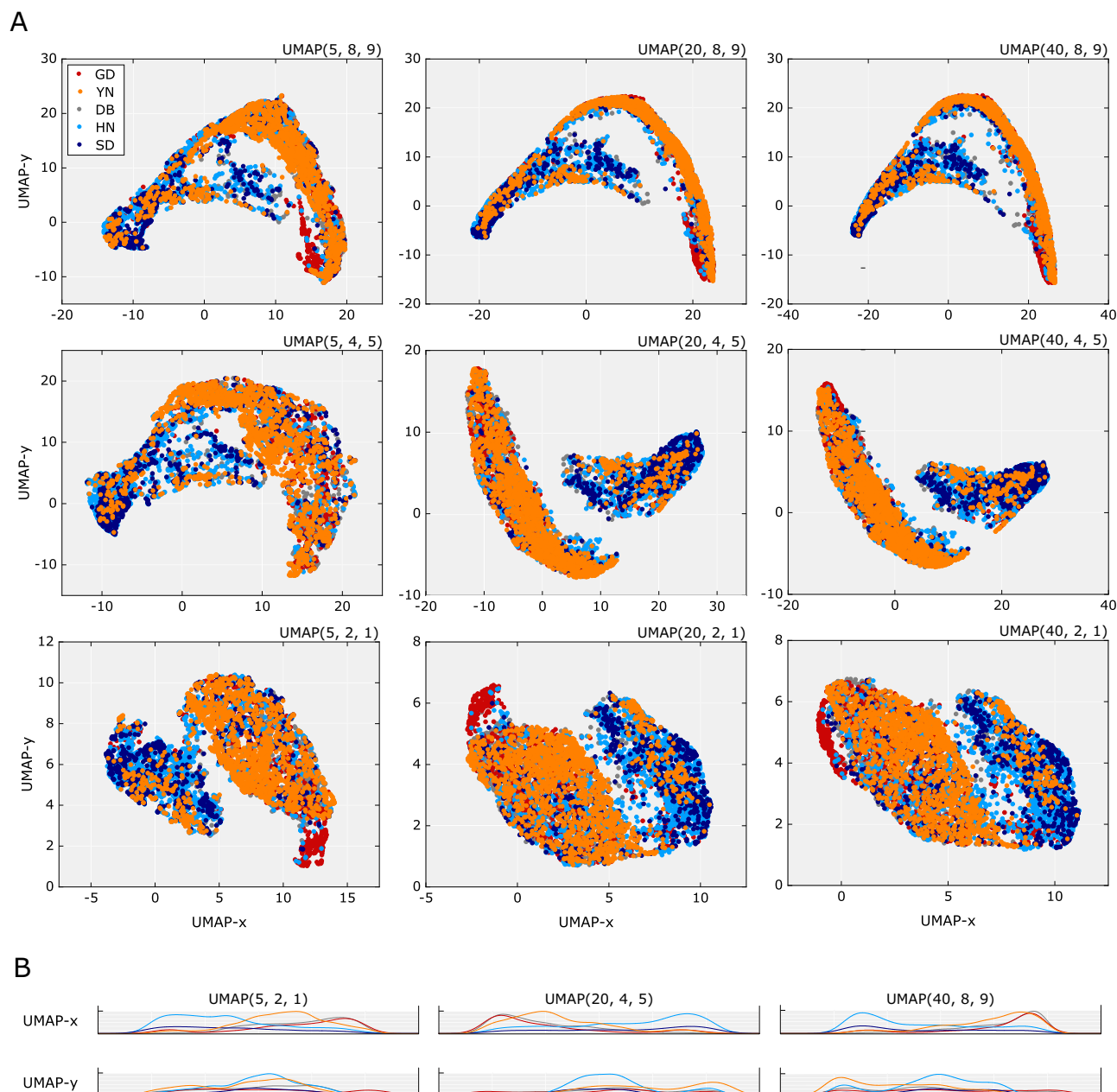


Fig. 5 | UMAP-based molecular thumbprint derived from peanut pod geometry. **A** Molecular thumbprint constructed from digitized images of single peanut pods, processed using UMAP-based learning, showcasing the impact of UMAP parametric variations. Peanuts were sourced from various cultivar regions in China: Guangdong (red), Yunnan (orange), Henan (blue), Shandong (dark blue), and

Dongbei (grey) (details in Supplementary Fig. 1). **B** Cross-sectional distribution along the X-axis (top) and Y-axis (bottom) for UMAP configurations (5, 2, 1), (20, 4, 5), and (40, 8, 9). UMAP parameters ranged within (n_neighbors: 5–50, components: 2–10, spread: 1–10), denoted as UMAP (n_neighbors, components, spread).

and between length (L) and weight (W) (biophysical traits). In contrast, a negative correlation was found between T_1 relaxation and length (T_1 -L) (Supplementary Table 3).

The representation of diverse phenotypic traits (biophysical and biochemical) across peanut pods from various regions in China is illustrated in Fig. 7. Peanuts are known to be less sensitive to light intensity but highly responsive to temperature fluctuations, influenced by factors such as the differential length of day and night²⁷. These environmental factors can alter flower pod size and chemical composition, further shaping the observed phenotype-genotype relationships.

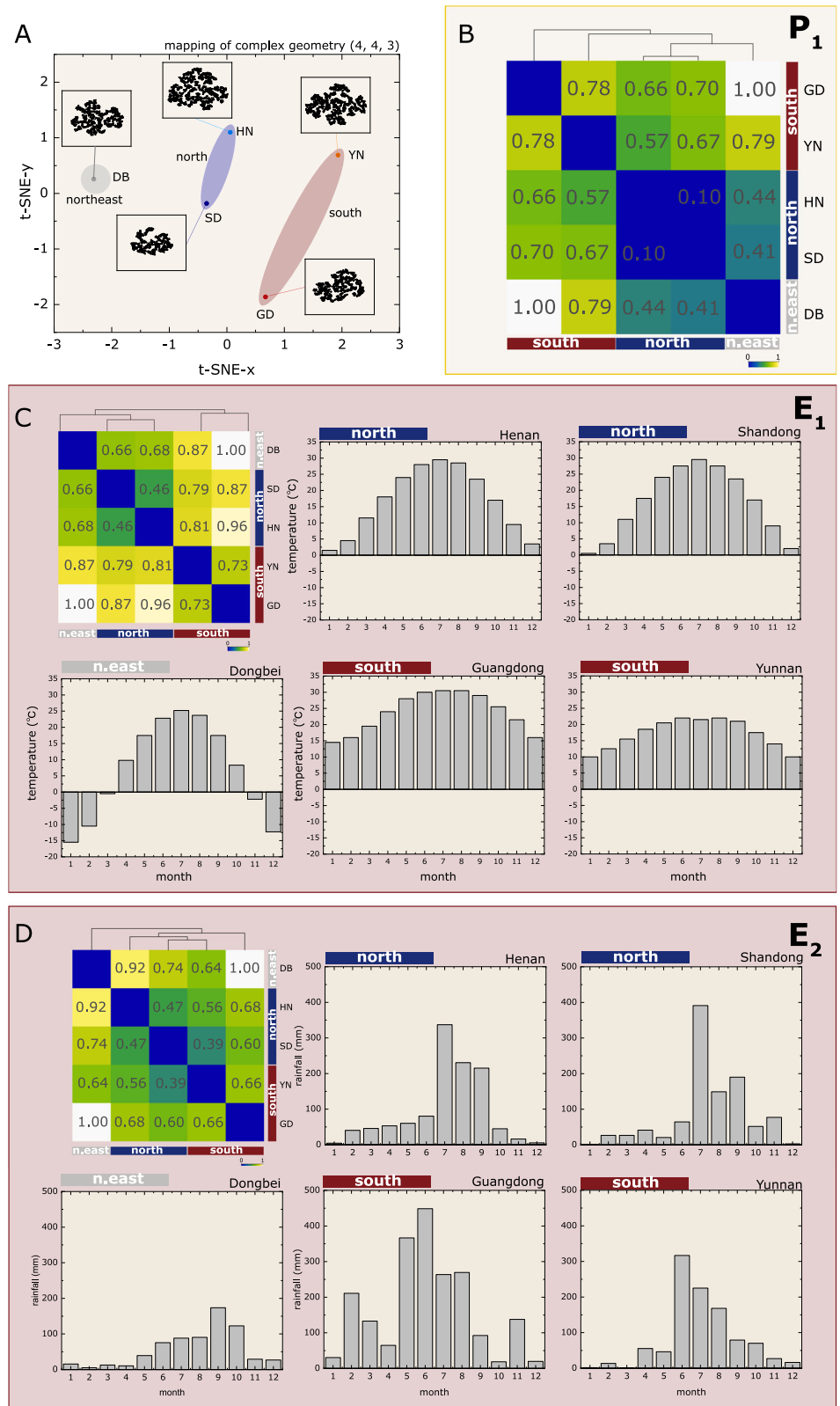
Discussion

This work establishes high-throughput phenotyping platforms of intricate geometric traits plants, harnessing the power of ubiquitous digital

technologies (such as smartphone, digital microscopy) integrated with the advanced data-driven approaches. While digitization ensures the preservation of dense information, enabling the creation and management of extensive databases, manifold learning facilitates the analysis of non-linear, high-dimensional datasets for complex geometry, a feat previously unachievable prior to the era of artificial intelligence. This framework, interestingly, can be a transformative approach to address the phenotyping bottleneck in plant breeding and far outreaching implication on human health and diseases.

Firstly, understanding the intricate geometry of peanut pods, which is a product of complex genotype-environment (GE) interactions is crucial for advancing the field phenomics^{28–31}, a key priority for agronomic communities in facing challenges from climate change, geopolitical instability, population growth, and depleting natural resources. Enhanced GE-space

Fig. 6 | Environmental Modulation of Geographically Distinct Molecular Phenomics Thumbprints in Peanut Pods. **A** Experimentally derived molecular thumbprints of peanut pods from different regions in China, analyzed using hierarchical clustering and t-SNE (perplexity=4, exaggeration=4, PCA components=3; details in Supplementary Fig. 3b). Regions are color-coded: south (red), north (dark blue), and northeast (grey). **B** Composite effect of the phenotypic variations, **C** Impact of annual temperature patterns and **D** rainfall patterns on molecular thumbprints, assessed via hierarchical clustering analysis (details in Methods). Data were collected in 2021.



insights enable the development of crops with higher yields, improved nutritional quality, and greater stress tolerance^{11,29,32}. These advancements are essential to meet escalating global food demands while ensuring sustainability and resilience in agriculture (Fig. 8).

Secondly, this study pioneered the identification of a geographically unique ‘molecular thumbprint’ for peanut cultivars, derived from their

topological signatures. This ‘molecular barcoding’ provides a robust, inherent authentication mechanism to counter food adulteration, a pressing challenge in ensuring food safety. By enabling precise traceability, this approach minimizes the risk of distributing unsafe or inferior products, thereby reducing product recalls. Moreover, enhanced traceability and transparency across the food supply chain promote fair competition and

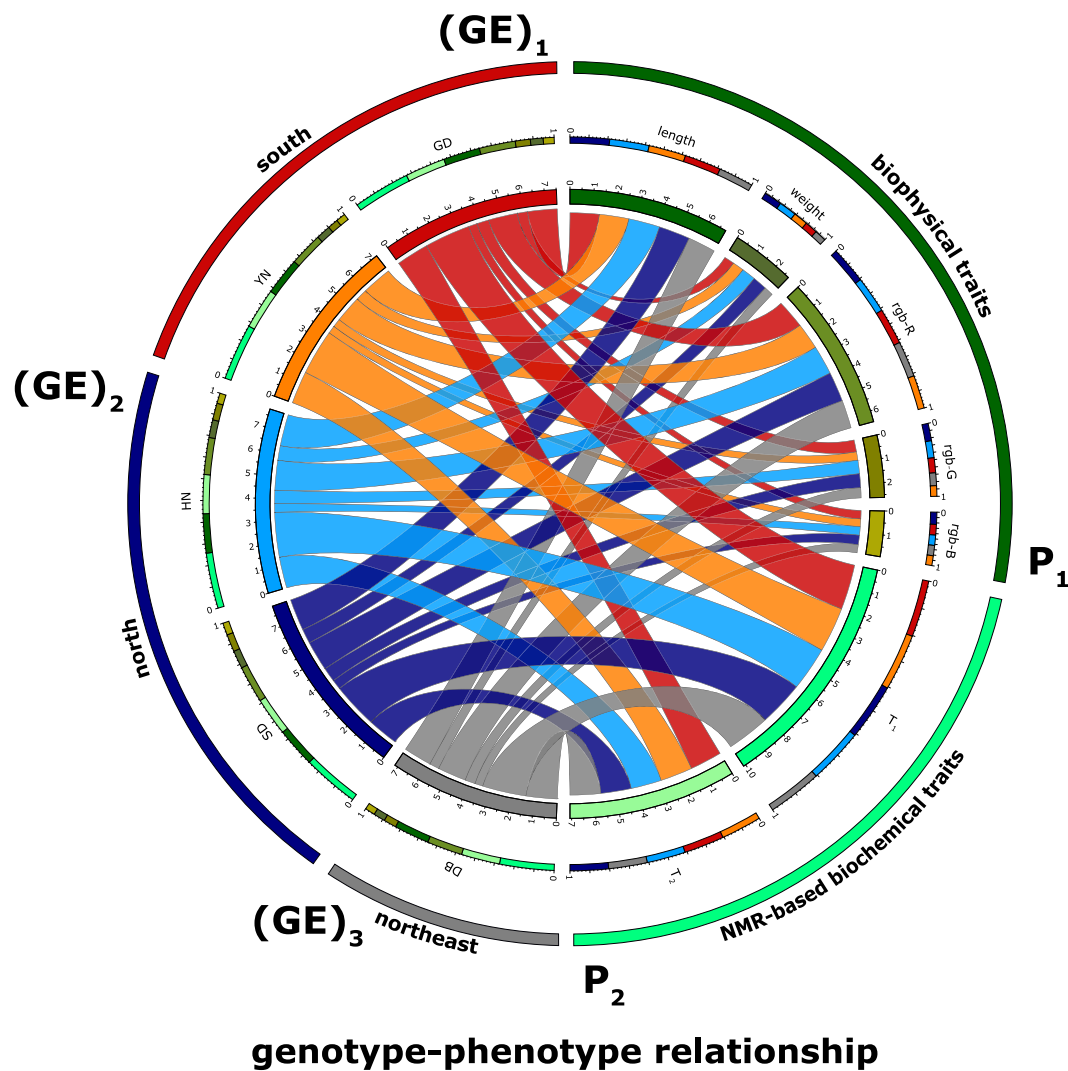


Fig. 7 | Circos representation of genotype–environment–phenotype relationships in peanut. This diagram summarizes the integrated associations among regional genotype–environment groups (GE)₁–(GE)₃ (outer colored arcs), key biophysical traits (P₁; top-right green sector), and NMR-based biochemical traits (P₂; bottom-right green sector). Each outer arc color denotes a geographic production region (red: South; blue: North; grey: Northeast). Inner radial tracks display trait distributions, where individual segments represent scaled trait values (e.g., pod length, height, geometric diameter, NMR-derived relaxation metrics, and RGB

intensity features). Chord ribbons encode the strength and direction of genotypic–phenotypic associations, with ribbon width proportional to relative contribution and ribbon color corresponding to the originating GE group. This visualization highlights region-specific morphological and biochemical signatures and illustrates how multi-dimensional phenotypes map onto distinct genotype–environment clusters within the proposed high-throughput manifold-learning framework.

bolster consumer trust by safeguarding against fraudulent practices^{21,33,34}. Our findings lay a foundation for scalable, innovative solutions in precision agriculture and food security.

Interestingly, the high-throughput phenotyping framework of intricate geometry we developed for plant sciences is readily applicable to human health and disease studies. The completion of the Human Genome Project in 2003, coincidentally augur the arrival of big data era (e.g., AI-driven analysis, and later the Large Language Model, LLM)^{35–37} and the development of new cutting-edge technologies (e.g., advanced imaging^{37–40}), which naturally forms a strong driving force to a new successor known as the Human Phenome Project (Fig. 9A). By leveraging phenomics, this approach unlocks new insights into human health and disease, offering perspectives unattainable through conventional genomic methods (Table 1)^{31,41–43}.

Diabetes, for example, is a multifaceted disorder modulated by numerous genes and external factors^{44–46}, blurred the boundary between “healthy” and “disease” states due to partially, its heterogeneity and inter-

individual variability (e.g., body composition, adiposity levels)^{45–48}. This high level of complexity demands cross-scale, multi-level analyses, driving the development of a novel clinical framework based on the clinical phenome (Fig. 9B, C)^{49,50}. Ironically, leveraging on the framework of intricate geometry phenotyping as proposed; such as captures intricate biophysical traits (such as intricate abdominal fat composition), may one day enables the prediction of diabetes-related phenotypes (e.g., insulin resistance, liver cirrhosis, gut microbiome)^{45–48} despite inter-individual variability. This approach offers a robust tool for precision phenotyping, enhancing our ability to understand and manage diabetes effectively.

The process of digitization facilitates preservation of high density of information, making the creation and management of large databases possible, making this proposal a comprehensive blueprint towards the development of a Large Geometric Model (LGM), analogous to Large Language Models (LLMs)^{51,52}, to capture complex geometric traits. This framework holds promise for near-real-time prediction, understanding, and

response to disease progression, akin to functional predictions in structural proteins using Large Protein Models (LPMs)⁵³.

Despite the strengths of this study, several limitations are acknowledged. First, the inherent “black-box” nature of manifold-based and deep-learning models may introduce latent biases that are not immediately observable yet could influence phenotypic interpretation, particularly in high-stakes agricultural decision-making. Although extensive internal validations and stability assessments were conducted (Supplementary Information), and results remained consistent across repeated runs, hyperparameter settings, and embedding

initializations, the possibility of unrecognized model behaviour cannot be fully excluded.

Second, the geographic composition of the dataset may introduce sampling biases. Regional over-representation has the potential to impose unintended topological constraints on the learned manifold, producing clusters driven by sampling structure rather than true phenotypic divergence. Such effects may manifest as spurious separations in t-SNE or UMAP space. Future work will incorporate stratified accession sampling, balanced augmentation strategies, and domain-adaptation layers to disentangle intrinsic pod geometry from geographic or environmental influences.

Third, although the current framework is demonstrated using peanut pods, the approach is inherently crop-agnostic. Because the manifold is derived directly from pixel-level geometric features, the same analytical pipeline can be applied to other crops such as soybean pods, barley spikes, and rice panicles without modification to the core architecture. Only the training image set needs to change. Ongoing studies on additional crops (e.g., potatoes, rice grains) are currently underway; with results to be reported in future work.

Finally, the development of a robust Large Geometric Model requires substantial quantities of high-quality, well-labelled images. Advances in computational parallelism (e.g., CUDA-enabled GPUs), modern neural architectures (e.g., Transformers), and minimally supervised learning have considerably improved the feasibility of constructing such models, yet the fundamental challenge of securing sufficient high-quality data remains.

In conclusion, as gene sequencing becomes increasingly accessible and cost-effective, the central challenge in the life sciences is shifting from generating genomic data to achieving comprehensive, high-resolution phenotypic measurement. With the expanding use of big-data and multi-dimensional strategies such as trans-omics, manifold-based learning is becoming essential for interpreting complex genotype–phenotype relationships. Our study demonstrates the feasibility of a high-throughput phenotyping framework grounded in manifold learning, signaling meaningful progress toward intelligent phenotyping. This work represents a transition from traditional genomics-centric approaches and lays the

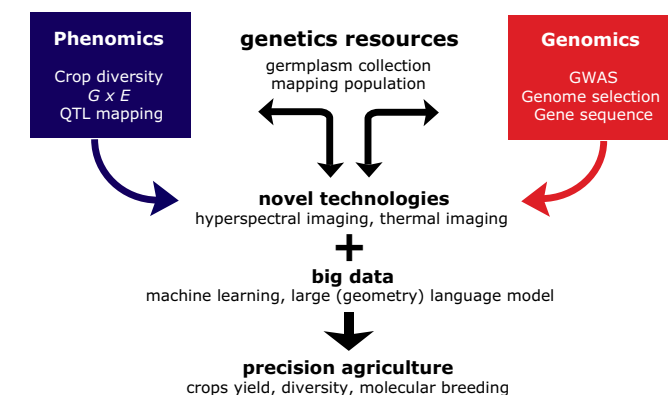


Fig. 8 | A paradigm shift in precision agriculture driven by advancements in gene-sequencing technologies, big data analytics (e.g., data-driven research, large language models), and innovative phenotyping technologies. (a) Integration of genomics and phenomics through leveraging genetic resources in the genomics space (e.g., genome-wide association studies [GWAS], genomic selection, gene sequencing) and phenomics space (e.g., exploration of crop diversity, genotype-by-environment [GxE] interactions, quantitative trait loci [QTL] mapping).

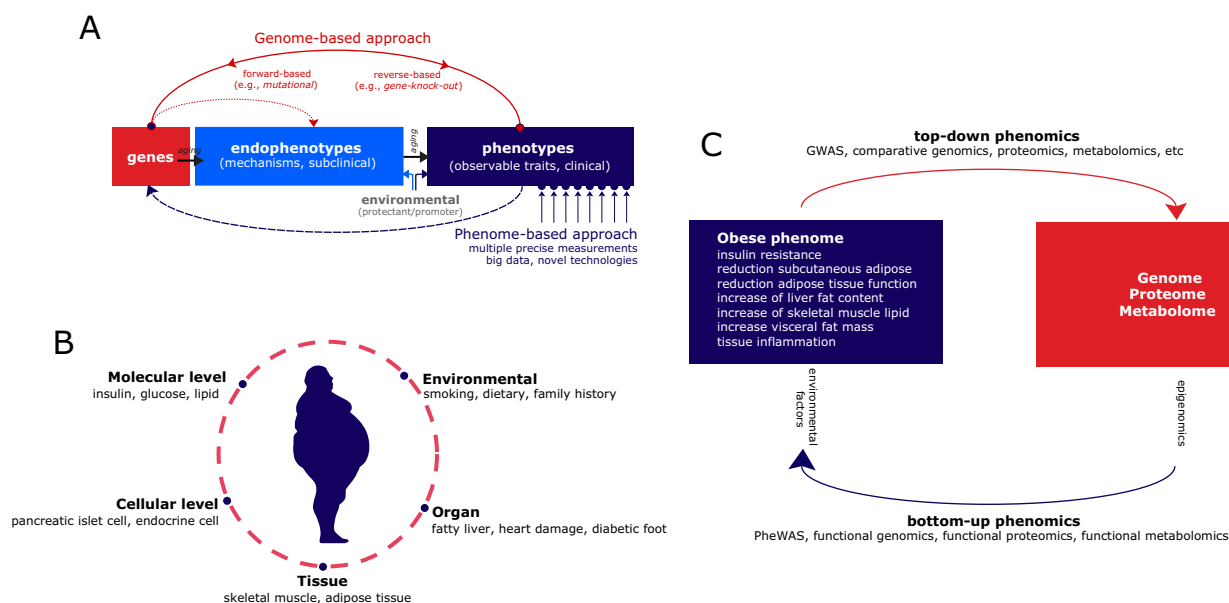


Fig. 9 | A paradigm shift in human health and disease research driven by multi-phenotyping technologies and large-scale data-driven phenotypic trait foundation models. **A** Traditional genomics-based approaches employ two strategies: forward genetics, which identifies phenotypes from naturally occurring mutations followed by genomic mapping for validation, and reverse genetics, using molecular techniques (e.g., CRISPR) to analyse gene function by altering nucleic acid sequences and observing phenotypic changes. In contrast, phenome-based research leverages advances in precision measurement technologies and data-driven foundation

models to capture comprehensive phenotypic data. **B** Chronic diseases like diabetes exhibit multi-level, cross-scale phenotypes, necessitating continuous integration of dynamic clinical observations. **C** The obese phenotype encompasses not only clinical metrics (e.g., elevated body mass index, waist circumference, body fat percentage) but also systematically defined phenotypic traits across molecular levels, including metabolites, proteins, and genes. It combines top-down approaches (e.g., GWAS, comparative genomics) with bottom-up validation (e.g., PheWAS, functional genomics), mirroring strategies from traditional genomics.

foundation for transformative applications in precision agriculture and medical phenomics.

Methods

Materials

A total of 6500 peanut pod samples were procured from five regions in China—Guangdong (GD: 1,069 samples), Yunnan (YN: 1,651), Hunan (HN: 1,786), Shandong (SD: 717), and Dongbei (DB: 1,201)—via the Taobao platform from 25 merchants (4 GD, 6 YN, 7 HN, 3 SD, 5 DB). Samples were labeled with regional abbreviations (e.g., GD1, GD2) to maintain merchant privacy, omitting commercial brand names. Samples were analyzed without further processing and stored in sealed Ziplock bags in a desiccator when not in use (details in Supplementary Fig. 1).

Sample Size Determination and Power Analysis

An a priori statistical power analysis was conducted before data collection to determine the minimum sample size required to robustly test the hypothesis that peanut pod morphology varies systematically across cultivation locations, thereby minimizing Type II error risk. The analysis was performed in G*Power software (Version 3.1.9.7) using a two-sample MANOVA framework (Global effect) on the embedding vectors, we computed required sample sizes for achieving 80% statistical power at $\alpha = 0.05$. The final dataset size were much larger ($N > \text{several hundred}$) than the obtained estimates (medium effect, N of 50 per region) (details in Supplementary Information).

Image Acquisition and Preprocessing

A low-cost microscope with an integrated camera (Shengchuang 1080 FHD, China), featuring a 6 cm focal length and 1,200x magnification with image resolution (2400×1800 pixels) was used to capture images of individual peanut pods (details in Supplementary Fig. 9). To eliminate stray light, a detector chamber with LED lighting (AMBITFUL, China) and an enclosure was employed. A digital light meter (UT383S UNI-T, China) standardized LED luminosity. Strict calibration addressed potential errors from digitization, including lighting bias, microscope limitations, image reproducibility, and subjective judgment (details in Supplementary Fig. 4).

Biophysical Traits Measurements

Of the 6,500 samples, 6,424 were sourced from 25 suppliers to capture a broad spectrum of peanut varieties cultivated across China (details in Supplementary Fig. 1). Biophysical traits, including length, weight, and size, were measured for each peanut pod using a Digimatic Caliper (San Liang Corporation, China), a high-precision electronic scale (Lea Ping, China), and nitrile examination gloves (INTCO, China).

NMR Measurements and Parameters

Proton (^1H) magnetic resonance measurements of peanut oil were conducted at a resonance frequency of 21.7579 MHz using a portable permanent magnet (Metrolab Instruments, Switzerland, $B_0 = 0.5$ T) and a benchtop console (Kea Magritek, New Zealand). The measurement chamber was maintained at 30 °C using a temperature controller. T_1 and T_2 relaxation times were measured using standard inversion recovery (IR) and Carr-Purcell-Meiboom-Gill (CPMG) pulse sequences, respectively. Experimental parameters included an echo time of 100 μs , 10,000 echoes, and 32 signal averages. A 2-second recycle delay ensured molecular spins returned to thermal equilibrium. Measurements were performed on oil extracted from single peanuts using an in-house cold-pressing method.

Material Preparations for Micro-NMR Analysis

Cold-pressed (CP) oil extraction was performed using a homemade mechanical presser (e.g., G-clamp) at room temperature. Each extraction involved a single peanut pod, with the resulting oil filtered through nanopore-like tissue papers to ensure purity.

Manifold Learning: t-SNE and UMAP Analysis

The goal of dimensionality reduction is to preserve the significant structure of high-dimensional data in a low-dimensional representation, typically a 2D or 3D map. Techniques such as t-distributed Stochastic Neighbor Embedding (t-SNE), Maximum Variance Unfolding (MVU), and Laplacian Eigenmaps were employed. However, most methods struggle to retain both local and global data structures in a single map.

t-SNE visualizes high-dimensional data by assigning each data point a position in a low-dimensional space. It calculates neighborhood information in the high-dimensional space (i.e., surrounding points for each data point) and maps this to the low-dimensional space. The Kullback-Leibler (KL) divergence is used to measure the difference between high- and low-dimensional representations, optimizing the low-dimensional map to minimize this divergence.

These techniques primarily serve as visualization tools, leaving data interpretation to the observer. A key parameter in t-SNE, perplexity, acts as a smooth measure of the effective number of neighbors, controlling the Gaussian kernel's width used to compute point similarities. Perplexity governs how many nearest neighbors each point is attracted to, with typical values ranging from 5 to 50 (default values often set at 30 or 50). The performance of t-SNE is generally robust to changes in perplexity, allowing flexibility in its application.

Mathematical Modeling of Complex Geometry of a Peanut Pod

To simulate the clustering effects of inter- and intra-species variations in peanut pods, two mathematical models were developed: Model-1 (size-based) and Model-2 (shape-based). Model-1 utilized third-order functions derived from the two-dimensional outline of peanut pods (Eq. 1). The Desmos mathematical function tool, RevolutionPlot3D (based on a revolution's cylindrical coordinates system), and Mathematica Wolfram software (built-in toolkit) were employed. Three-dimensional simulation models were generated by rotating the extracted function once, capturing size-based variations. Model-2 was developed similarly, using equations (Eqs. 2–6) to represent shape-based variations. Like Model-1, it generated three-dimensional models to reflect shape-driven differences.

The resulting models (images of peanut pods) were processed using statistical programming languages, specifically Orange 3.1.2. Images were converted into image embeddings using Google's Inception v3 model, pre-trained on ImageNet. Machine learning techniques, including t-SNE, Multidimensional Scaling (MDS), and Isomap, were applied for dimension reduction. Manifold-based processing was completed in less than 1 minute on a desktop computer equipped with a 12th Gen Intel® Core™ i9-12900K CPU @ 3.19 GHz.

The size-based Model-1 and shape-based Model-2 effectively simulated inter- and intra-species clustering of peanut pods. The use of advanced tools and machine learning techniques enabled rapid processing and visualization of complex geometric traits, supporting the study's analysis of phenotypic diversity.

$$F(x) = \begin{cases} R - x^3 e^{0.1x}, & x \geq 0 \\ R + x^3 e^{-x}, & x < 0 \end{cases} \quad (1)$$

$$F(x) = \begin{cases} R - x^3 e^x, & x \geq 0 \\ R + x^3 e^{-x}, & x < 0 \end{cases} \quad (2)$$

$$F(x) = R - x^4 e^{-x} \quad (3)$$

$$F(x) = \begin{cases} R - x^2 \sin x, & x \geq 0 \\ R + x^{10} \sin x, & x < 0 \end{cases} \quad (4)$$

$$F(x) = R - x^8 e^{-x} \quad (5)$$

$$F(x) = \sqrt{\frac{1}{a^2}(a^2b^2 - b^2x^2)} \quad (6)$$

Machine Learning Algorithm and Workflows

Raw datasets were processed using unsupervised learning techniques implemented in a statistical programming language, specifically Orange 3.32.0 or R. The machine learning algorithms were executed on a personal laptop equipped with an Intel Core Pentium i5 CPU @ 2.40 GHz and 8.00 GB RAM. Once the machine learning model was constructed, all tasks ran concurrently, typically completing in less than 1 minute.

Clustering analysis, including tree classification and hierarchical clustering (HC), was used to rapidly establish relationships between objects, with quantitative linkages (e.g., inter- and intra-cluster similarity) visualized through dendrograms and heatmaps. For data exploration and visualization of high-dimensional data (e.g., UV-Vis spectrum, wavelength range 300–550 nm), an unsupervised non-linear dimensionality reduction technique, t-Distributed Stochastic Neighbor Embedding (t-SNE), was employed. The t-SNE parameters were configured as follows: Perplexity (1–500), Exaggeration (1–4), and PCA components (2–50).

Deep learning: Image Embedding and Cross Validation

To substantiate the hypothesis of geographically unique molecular thumbprints encoded in peanut pod geometries, a visual analytics workflow was developed in Orange Data Mining. This process began with the importation of 6,500 pod images from diverse geographic origins using the Import Images widget, where subfolder structures automatically designated location-specific class labels. Subsequent transformation employed the Image Embedding widget, leveraging pre-trained convolution neural networks such as InceptionV3 or ResNet50 to convert each image into a 2048-dimensional feature vector. These vectors captured nuanced morphological variations potentially reflective of underlying molecular profiles, and the resulting embeddings were organized into a Data Table for further analysis, establishing a foundation for rigorous quantitative evaluation of geographic distinctiveness.

Statistical robustness was ensured through comprehensive validation techniques, including 10-fold cross-validation and leave-one-out cross-validation within the Test & Score widget, applied to the embeddings using classifiers like Random Forest, Support Vector Machines, and k-Nearest Neighbours. Performance metrics, such as mean accuracy and per-class F1-scores, underscored the reliable predictive separation of origins (details in Supplementary Fig. 10).

Statistics and reproducibility

All statistical analyses are described in the Methods section, with significance defined as $p < 0.05$ (unless mentioned otherwise). Experimental included at least three replicates, and results are reported as mean $\pm$ SD (median as stated). To compare the means of two independent groups (e.g., NMR-based traits), Student's t-test was used. The hypothesis that peanut pod geometry reflects unique regional molecular signatures was tested using multivariate statistics combined with manifold learning. Cultivation region (South; North; Northeast) was examined as the factor and the four morphometric traits as outcomes in a one-way MANOVA (see Methods, Supplementary Information). Region-specific effects were pinpointed through follow-up univariate ANOVAs, trait links were measured via Pearson correlations (with p-values and 95% CIs). Statistical software used were G*Power 3.1.9.7, Orange Data Mining (3.39.0), and OriginPro8.

Data availability

The image dataset of peanut pods analyzed in this study and the extracted phenotypic trait data are publicly available in the GitHub repository: <https://github.com/pengwengkung/Complex-geometry-peanut>.

Code availability

The Orange Data Mining workflow file (.ows) used for analysis is available in the same GitHub repository: <https://github.com/pengwengkung/Complex-geometry-peanut>.

Received: 19 May 2025; Accepted: 7 March 2026;

Published online: 19 March 2026

References

- Wanget, S. et al. Agro-morphological and chemical characterization of traditional Indonesian peanut (*Arachis hypogaea* L.) cultivars. *Genetika*. **51**, 179–198 (2019).
- Zhuang, W. et al. The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nat. Genet.* **51**, 865–876 (2019).
- Ren, X. et al. Genetic Diversity and Population Structure of the Major Peanut (*Arachis hypogaea* L.) Cultivars Grown in China by SSR Markers. *PLoS ONE* **9**, e88091 (2014).
- Li, Y., Zhang, R., Qin, X., Liao, Y. & Siddique, K. H. M. Changes in the protein and fat contents of peanut (*Arachis hypogaea* L.) cultivars released in China in the last 60 years. *Plant Breed.* **137**, 746–756 (2018).
- Großkinsky, D. K., Svendsgaard, J., Christensen, S. & Roitsch, T. Plant phenomics and the need for physiological phenotyping across scales to narrow the genotype-to-phenotype knowledge gap. *J. Exp. Bot.* **66**, 5429–5440 (2015).
- Peng, W. K., Chen, L., Boehm, B. O., Han, J. & Loh, T. P. Molecular phenotyping of oxidative stress in diabetes mellitus with point-of-care NMR system. *Npj Aging Mech. Dis.* **6**, 11 (2020).
- Ren, A. et al. Evaluation of an intelligent artificial climate chamber for high-throughput crop phenotyping in wheat. *Plant Methods* **18**, 77 (2022).
- Tiwari, J. K. et al. High-Throughput Phenotyping in Potato Breeding. in *High-Throughput Crop Phenotyping* (eds. Zhou, J. & Nguyen, H. T.) 165–182 (Springer International Publishing, Cham, 2021) https://doi.org/10.1007/978-3-030-73734-4_8.
- Peng, W. K., Ng, T.-T. & Loh, T. P. Machine learning assistive rapid, label-free molecular phenotyping of blood with two-dimensional NMR correlational spectroscopy. *Commun. Biol.* **3**, 535 (2020).
- Slater, A. T. et al. Breeding differently—the digital revolution: high-throughput phenotyping and genotyping. *Potato Res.* **60**, 337–352 (2017).
- Ubbens, J. R. & Stavness, I. Deep plant phenomics: a deep learning platform for complex plant phenotyping tasks. *Front. Plant Sci.* **8**, 1190 (2017).
- Yurkovich, J. T., Tian, Q., Price, N. D. & Hood, L. A systems approach to clinical oncology uses deep phenotyping to deliver personalized care. *Nat. Rev. Clin. Oncol.* **17**, 183–194 (2020).
- Houle, D., Govindaraju, D. R. & Omholt, S. Phenomics: the next challenge. *Nat. Rev. Genet.* **11**, 855–866 (2010).
- Nawade, B. et al. Insights into the Indian Peanut Genotypes for ahFAD2 Gene Polymorphism Regulating Its Oleic and Linoleic Acid Fluxes. *Front. Plant Sci.* **7**, (2016).
- Ali, S. et al. Groundnut genotypes' diversity assessment for yield and oil quality traits through multivariate analysis. *SABRAO J. Breed. Genet.* **54**, 565–573 (2022).
- Bertioli, D. J. et al. The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nat. Genet.* **48**, 438–446 (2016).
- D'Orazio, M. et al. Machine learning phenomics (MLP) combining deep learning with time-lapse-microscopy for monitoring colorectal adenocarcinoma cells gene expression and drug-response. *Sci. Rep.* **12**, 8545 (2022).
- Peng, W. K., Chen, L. & Han, J. Development of miniaturized, portable magnetic resonance relaxometry system for point-of-care medical diagnosis. *Rev. Sci. Instrum.* **83**, 095115 (2012).

19. Peng, W. K. et al. Micromagnetic resonance relaxometry for rapid label-free malaria diagnosis. *Nat. Med.* **20**, 1069–1073 (2014).
20. Peng, W. K. Clustering nuclear magnetic resonance: machine learning assistive rapid two-dimensional relaxometry mapping. *Eng. Rep.* **3**, (2021).
21. dos Santos, V. R. et al. Novel time-domain NMR-based traits for rapid, label-free Olive oils profiling. *Npj Sci. Food* **6**, 59 (2022).
22. Zimmer, A. et al. The geometry of clinical labs and wellness states from deeply phenotyped humans. *Nat. Commun.* **12**, 3578 (2021).
23. Roca, C. P. et al. A cross entropy test allows quantitative statistical comparison of t-SNE and UMAP representations. *Cell Rep. Methods* **3**, 100390 (2023).
24. Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* **10**, 5416 (2019).
25. Parent, B. & Tardieu, F. Can current crop models be used in the phenotyping era for predicting the genetic variability of yield of plants subjected to drought or high temperature? *J Exp. Bot.* **65**, 6179–6189 (2014).
26. Shang, L. et al. A super pan-genomic landscape of rice. *Cell Res.* **32**, 878–896 (2022).
27. Rowell, T., Mortley, D. G., Loretan, P. A., Bonsi, C. K. & Hill, W. A. Continuous daily light period and temperature influence peanut yield in nutrient film technique. *Crop Sci.* **39**, 1111–1114 (1999).
28. Wagner, G. P. & Zhang, J. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. *Nat. Rev. Genet.* **12**, 204–213 (2011).
29. Gratani, L. Plant phenotypic plasticity in response to environmental factors. *Adv. Bot.* **2014**, 1–17 (2014).
30. Virolainen, S. J., VonHandorf, A., Viel, K. C. M. F., Weirauch, M. T. & Kottyan, L. C. Gene–environment interactions and their impact on human health. *Genes Immun.* **24**, 1–11 (2022).
31. Jiang, N. & Zhu, X.-G. Modern phenomics to empower holistic crop science, agronomy, and breeding research. *J. Genet. Genomics.* **51**, 790–800 (2024).
32. Lv, Z. et al. The determination of peanut (*Arachis hypogaea* L.) pod-sizes during the rapid-growth stage by phytohormones. *BMC Plant Biol.* **23**, 371 (2023).
33. Pirs, S., Banafshechin, E., Amiri, S., Rahimirad, A. & Ghafarzadeh, J. Detection of fraud of palm, sunflower, and corn oil in butter using HPLC profile of tocopherols and tocotrienols by response surface method. *J. Iran. Chem. Soc.* **18**, 1167–1177 (2021).
34. Azizian, H. et al. Novel, rapid identification, and quantification of adulterants in extra virgin olive oil using near-infrared spectroscopy and chemometrics. *Lipids* **50**, 705–718 (2015).
35. Halama, A. et al. A roadmap to the molecular human linking multiomics with population traits and diabetes subtypes. *Nat. Commun.* **15**, 7111 (2024).
36. Freimer, N. & Sabatti, C. The Human Phenome Project. *Nat. Genet.* **34**, 15–21 (2003).
37. Qi, Y.-J. et al. Radiomics in breast cancer: current advances and future directions. *Cell Rep. Med.* **5**, 101719 (2024).
38. Weissleder, R. & Nahrendorf, M. Advancing biomedical imaging. *Proc. Natl. Acad. Sci. USA* **112**, 14424–14428 (2015).
39. Weissleder, R. & Pittet, M. J. Imaging in the era of molecular oncology. *Nature* **452**, 580–589 (2008).
40. Parekh, V. & Jacobs, M. A. Radiomics: a new application from established techniques. *Expert Rev. Precis. Med. Drug Dev.* **1**, 207–226 (2016).
41. Esposito, S., Carputo, D., Cardi, T. & Tripodi, P. Applications and trends of machine learning in genomics and phenomics for next-generation breeding. *Plants* **9**, 34 (2019).
42. Li, H. et al. An integrated systems genetics and omics toolkit to probe gene function. *Cell Syst.* **6**, 90–102.e4 (2018).
43. Tian, M. et al. Report on the 3rd Board Meeting of the International Human Phenome Consortium. *Phenomics* **3**, 77–82 (2023).
44. Yugi, K. et al. Reconstruction of insulin signal flow from phosphoproteome and metabolome data. *Cell Rep.* **8**, 1171–1183 (2014).
45. Peng, W. et al. Metabolite profiles of distinct obesity phenotypes integrating impacts of altitude and their association with diet and metabolic disorders in Tibetans. *Sci. Total Environ.* **949**, 174754 (2024).
46. Li, L. et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci. Transl. Med.* **7**, (2015).
47. Coral, D. E. et al. A phenome-wide comparative analysis of genetic discordance between obesity and type 2 diabetes. *Nat. Metab.* **5**, 237–247 (2023).
48. Zhang, Y.-P., Zhang, Y.-Y. & Duan, D. D. From Genome-Wide Association Study to Phenome-Wide Association Study: New Paradigms in Obesity Research. in *Progress in Molecular Biology and Translational Science* vol. 140 185–231 (Elsevier, 2016).
49. Piché, M.-E., Tchernof, A. & Després, J.-P. Obesity Phenotypes, Diabetes, and Cardiovascular Diseases. *Circ. Res.* **126**, 1477–1500 (2020).
50. Han, Y. et al. Phenomics of Vascular Disease: The Systematic Approach to the Combination Therapy. *Curr. Vasc. Pharmacol.* **13**, 433–440 (2015).
51. Wang, J. et al. Large Language Model for Geometric Algebra: A Preliminary Attempt. in *Advances in Computer Graphics* (eds. Sheng, B., Bi, L., Kim, J., Magnenat-Thalmann, N. & Thalmann, D.) vol. 14498 237–249 (Springer Nature Switzerland, Cham, 2024).
52. Yap, C.-K. Towards exact geometric computation. *Comput. Geom.* **7**, 3–23 (1997).
53. Madani, A. et al. Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **41**, 1099–1106 (2023).

Acknowledgements

This research was supported by Quantum Information and Quantum Biology, The University of Osaka, and the Frontier Research Centre of Songshan Lake Materials Laboratory. The funders are greatly acknowledged - Frontier Research Centre of Songshan Lake Materials Laboratory (Y1D111M511), Research Fund for International Excellent Young Scientists (T2350610282), Guangdong Basic and Applied Basic Research Foundation, Guangdong Dongguan Joint Fund Project, Regional Incubation Project 2023, Natural Science Foundation of Guangdong Province, General Program 2024.

Author contributions

Weng Kung Peng: Conceptualization, Formal analysis, Writing – original, review & editing, Supervision, Funding acquisition, Xiaomin Lin : Writing – review & editing, Formal analysis, Data curation. Chunling Ding, Peishan Deng: Formal analysis, Data curation. Jianwen Jiang: Simulation, Formal analysis, Data curation. Shijin Yuan: Software, Project administration.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41540-026-00688-1>.

Correspondence and requests for materials should be addressed to Weng Kung Peng.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026