

Research Article

RootMap: A Longevity Medicine Framework for Mapping Conditional Dependencies Between Aging Hallmarks and Organ-Aging Patterns

Jianghui Xiong¹

1. DeepoMe Inc., China

Longevity medicine needs ways to identify a small number of candidate intervention targets that might affect several functions at once. Protein–protein interaction (PPI) networks help map molecular connections and rank candidate drugs, but they do not show which maintenance states are associated with age-related patterns in particular organs. We introduce RootMap, a blood DNA-methylation framework for mapping these conditional dependencies, and apply it to 656 discovery and 1,394 replication participants. Each of 332 gene modules is scored as a module-level intrinsic capability (mIC). We define a conditional dependency as a difference in an organ module's association with age between participants with higher versus lower age-adjusted mIC for a hallmark module.

Twenty-two hallmark–organ pairs passed four artifact checks; 17 of 20 pairs on the replication list kept their direction in the second cohort, including all 10 testable organ–line pairs. The pattern was asymmetric: hallmark modules more often marked differences in organ–age associations than organs marked differences in hallmark–age associations. Stem-cell maintenance was the broadest marker in the discovery cohort, relating to 12 organ modules; senescence regulation was broadest in the replication cohort. The meaning of dependency is illustrated by bone marrow: among people with higher age-adjusted stem-cell-maintenance mIC, bone-marrow mIC was negatively associated with age ($r = -0.241$), while among those with lower mIC the association was close to zero ($r = +0.005$). The map also distinguishes organs by their patterns of dependencies: for example, the liver had several hallmark associations, including associations in opposite directions. These are patterns in blood-based scores, not proof of biological control.

The map may help drug screening by adding information about biological layer and dependency to molecular proximity. In a benchmark, filtering a protein–interaction (PPI)–ranked anti-aging

candidate list by functional-layer labels made the list 2.8 times more compact and increased hit enrichment from 3.19-fold to 8.87-fold. This gain was not distinguishable from draws matched for the number of drug targets ($p = 0.18$), and the filter used layer labels rather than learned dependency edges. Thus, the result supports testing layer information alongside PPI—not claiming that dependency edges have already improved drug discovery. Separately, discovery dependency effect sizes were associated with replication dependencies ($\rho = +0.153$, $p = 0.005$), whereas the tested PPI-proximity measures were not. Functional modules also showed intervention-associated changes more often than hallmark modules (25.4% versus 14.4% of tests); this measures response in the studied intervention series, not clinical benefit.

Traditional Chinese medicine (TCM) concepts were analyzed separately within the same framework. Essence was the largest hub, contributing 5 of 10 significant fundamental-substance pairs. The TCM concept “liver” corresponded more strongly to lymph and immune modules than to the anatomical liver ($r = -0.15$), illustrating that a traditional concept need not map onto the organ with the same name.

RootMap offers longevity medicine a framework for connecting foundational maintenance states with organ-aging patterns. By distinguishing capabilities associated with broad reach from functions that showed more frequent score changes in the intervention series, it gives researchers a practical basis for deciding what to measure, which candidate targets to investigate, and which functions to follow in longitudinal and N-of-1 studies. The dependency measures carried modest but reproducible information across cohorts beyond the PPI-proximity measures tested here, while layer-label filtering compacted a PPI-ranked candidate list in the screening benchmark. Together, these findings position RootMap as an evidence-graded map for prioritizing hypotheses in longevity research and as a foundation for future individual-level biomedical world models. Learned dependency edges and absolute mIC values remain priorities for further validation and calibration.

Corresponding author: Jianghui Xiong, jianghui@deepome.com

Introduction

Protein–protein interaction (PPI) networks have made molecular connectivity easier to map. They help researchers group genes and proteins into modules, connect modules with diseases, and search for anti-aging agents ^[1]. But a molecular interaction map is not yet a guide to intervention. It shows which

molecules are connected; it does not, by itself, show which biological processes sit upstream of particular organ functions, or which few intervention targets might influence several functions at once.

Aging-hallmark catalogs name processes shared across aging and age-related disease ^{[2][3]}. Yet they are usually treated as a list of mechanisms, not as a set of relationships between foundational processes and specific organ functions. At the same time, population-scale health world models use latent states to predict disease and treatment trajectories from large longitudinal datasets ^[4]. They address how health may change over time, but their predictive states do not necessarily identify which capabilities condition which functions. These approaches solve important problems, but leave open a practical question: can we measure the direction of functional dependencies well enough to help identify candidate intervention points?

We start from a possibility that is not established: biological capabilities may be organized in layers. Processes that maintain, repair, or renew cells and tissues could sit upstream of specialized functions such as organ, metabolic, and immune function. If so, a small number of foundational states might be related to age-associated patterns across several functions. This is a hypothesis to test, not an assumption that hallmarks are causal “root” mechanisms. In this paper, “upstream” refers to the proposed direction of a functional dependency; it does not mean that causality has been demonstrated.

To make the idea testable, we distinguish three levels: phenotypes (what presents clinically), specific functions (organ, metabolic, immune, and intrinsic-capacity modules), and foundational capabilities (maintenance, repair, renewal, and regulatory domains). We represent the foundational level with aging-hallmark-derived modules, retaining their original gene sets. This reframes the hallmarks as candidate capability domains and lets us ask how their measured states relate to specific functions, without treating the concepts as interchangeable.

This study extends our steerability framework for biomedical world models ^{[5][6]}. SteeraMed is a suite of biomedical world-model frameworks with modules for distinct tasks, including N-of-1 intervention reasoning and RootMap. Capomics (“omics of intrinsic capability”) measures module-level intrinsic capability (mIC), and RootMap maps conditional dependencies among those states. We plan to release a RootMap tool through steeramed.com. This paper presents and evaluates the RootMap module, not the full SteeraMed suite. Earlier work used a 332-module dictionary of human capability states to build virtual-patient models and prioritize drug-repurposing candidates from module panels ^[7]. Those analyses treated modules as separate units. Here, we apply Capomics to blood DNA methylation: each mIC

score is one minus the normalized module aging index, with bounds fixed in a reference cohort. Higher mIC indicates a higher estimated capability on the reference scale.

We define a dependency operationally. For a hallmark module and an organ module, we compare the organ module's association with age between people with relatively high and low age-adjusted mIC for the hallmark. A difference between those associations is a dependency in this analysis—not evidence that the hallmark causes organ aging. Because blood methylation is strongly shaped by cell composition, and because gene sets can overlap, we test the dependencies with controls for global cell-composition factors and shared gene content.

We apply this approach to two cohorts: 656 discovery participants and 1,394 replication participants. We examine whether the dependencies are sensitive to alternative gene-set definitions or to relearning the scoring procedure, and whether the same capability framework can be used to examine traditional Chinese medicine (TCM) concepts. We also use eight intervention series and separate transfer and drug-repositioning comparisons to ask what information the layered map might add to intervention research. These are tests of association, transfer, and candidate prioritization; they do not establish causal effects or clinical benefit. RootMap is a molecular dependency map, not yet a model that simulates a person's environment or response to treatment over time. For longevity medicine, its potential role is to help researchers decide which capability states and organ functions to measure or follow up in future intervention studies. It does not yet predict an individual's response or recommend a treatment.

Results

Scoring modules across the two medical systems

To build RootMap, we organized 332 predefined gene modules into a panoramic framework for studying aging across levels of function (Fig. 1a, b). Its foundation is 14 aging-hallmark modules (A1), treated as candidate capabilities that may support function. Three functional pillars represent organ aging (36 modules, A2), metabolic aging (10, A3), and immune aging (6, A4). Together, the foundation and these three pillars make up the 66-module modern system. Six additional intrinsic-capacity modules (A5) extend the map to broader whole-person function. A parallel TCM system contains 38 concepts: 7 fundamental substances such as qi and essence (T1), 13 TCM organs (T2), and 18 pathological patterns (T3). The remaining 222 modules—122 nutrient modules (NUT/NUTX) and 100 food-herb items (YFY)—

support phenotype mapping and candidate retrieval. The six A5 modules and these 222 modules are not included in the main comparisons.

Each module is scored in every blood sample. We first calculate the promoter-aggregated index (PAI) by comparing methylation in genes whose promoter methylation rises with age with genes whose methylation falls with age. These gene directions are selected in cohort A and then held fixed for cohort B. The analysis uses the resulting module-level intrinsic-capability score (mIC; Methods). The gene-level promoter-aggregation method follows our earlier cohort-mining work ^[8].

To test whether choosing age-related genes makes the module scores appear more reliable than they are, we repeated the scoring with five-fold cross-fitting. In each round, genes were selected using four-fifths of the participants, and the remaining fifth—unused in gene selection—was scored. All 18 organ-line modules kept the same direction of association with age across folds; 17 were significant in every fold (median $|r|$ attenuation = 0.96). The dependency edges were less stable: after reselecting genes and recalculating the scores, only 7 of the 14 pairs testable in cohort B kept the same direction. When genes were not reselected, all 14 pairs kept their direction. Thus, the module-level age associations were stable, but the dependency edges were sensitive to how the scoring genes were selected. We report score stability, edge stability, and network re-detection separately.

We ask whether a hallmark module changes how a function's score relates to age, rather than simply asking whether the two scores are correlated. To test this, we compare the function-age correlation in people with higher versus lower age-adjusted scores for the hallmark module. We call the difference between these two correlations Δz : a positive value means the association is more positive in the higher-score group; a negative value means it is more negative. The two correlations themselves (r_{high} and r_{low}) are provided for each pair in the versioned data package.

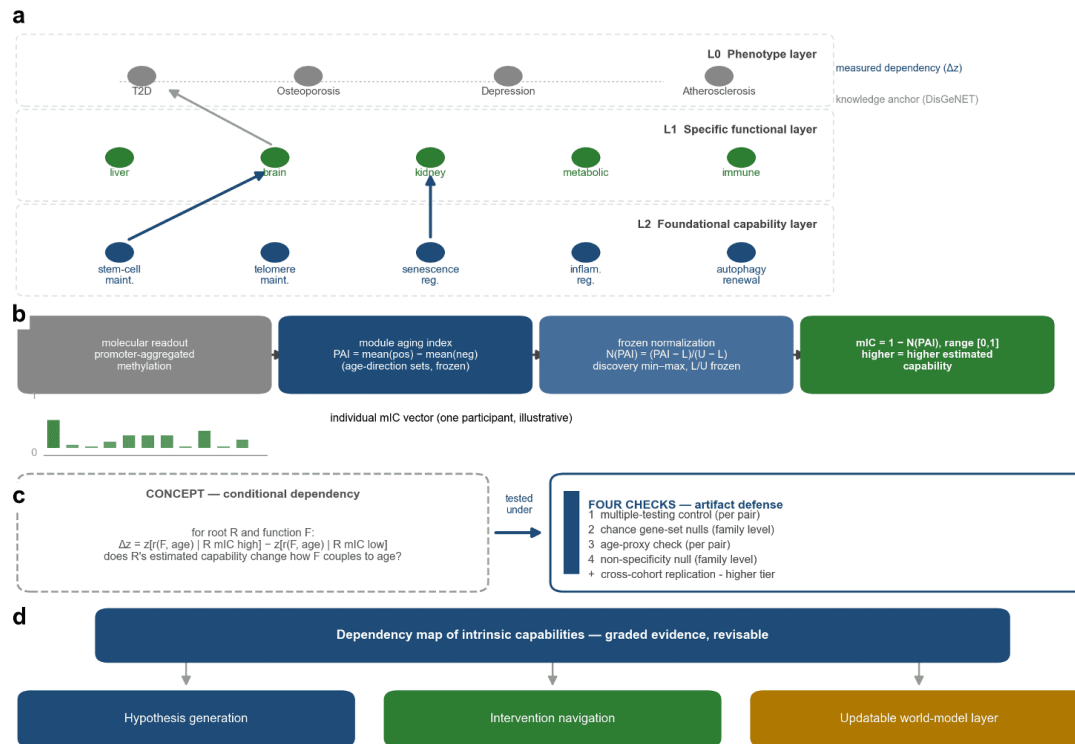


Figure 1. From clinical phenotype to function to foundational capability. (a) The map has three levels: phenotypes (clinical problems), specific functions (organ, metabolic and immune modules), and foundational capabilities (14 hallmark-derived modules). Gray dashed links show knowledge-based phenotype-to-module mappings, which were generated with computer assistance and still need expert review. Blue arrows show measured dependencies, not proven causal pathways. Table S10 lists each hallmark's source; Methods gives the full 332-module dictionary. (b) How mIC is calculated: blood methylation is summarized as a module aging index (PAI), scaled using bounds fixed in the reference cohort, then reversed so higher mIC means higher estimated capability: $mIC = 1 - N(PAI)$. Values fall between 0 and 1 in the reference cohort. In other cohorts, the same linear conversion is applied without cutting off values outside that range; this leaves Δz and paired-test results unchanged, but does not ensure that absolute mIC values are calibrated across cohorts (Table S13). The example vector shows several module scores for one person. (c) For each pair, we ask whether a function module's association with age differs between people with higher and lower scores for a hallmark module; the difference is Δz . Each claim must pass four artifact checks, with replication in a second cohort graded separately. (d) The map is intended to support hypothesis generation and intervention planning. These uses are proposed, not established causal pathways.

We assess the stability of individual relationships separately from the stability of the network's overall pattern. Seven pairs could not be tested because they involved two small modules—tendon (7 genes) and microbiota—that did not have enough selected genes whose methylation changes with age in the required directions in one or more cross-validation folds. This means there was not enough information to run the test for those pairs; it does not mean they were tested and found to have no dependency. The age-related gene lists selected in different folds were similar: the median Jaccard overlap was 0.73, meaning that the genes shared by two lists made up 73% of all distinct genes in either list.

To test whether the network's overall pattern could be reproduced after selecting genes and recalculating scores, we compared only pairs that could be tested in each analysis and matched the chance comparisons to those same pairs. First, we kept the original candidate pairs fixed. The link-concentration measure rose from 0.14 to 0.21 after cross-fitting, but the most connected regulator still had three links. The apparent rise came from testing fewer pairs (14 rather than 21), not from a regulator gaining links. Next, we searched again for dependencies among the testable pairs. This found more links, but did not recover the original links that passed all four checks more often than chance: across 20 repeats, 49 original links were recovered, compared with 54.6 expected by chance. A separate control scored each held-out group using a gene list from a different fold. Although this increased the number of detected links, it reused participants across the training and test groups; we therefore report it only as a secondary check and do not use it to explain the increase. Two pairs show why individual links need separate reporting: stem cell–bone marrow changed direction, whereas senescence–kidney kept its direction. In the discovery cohort, the main hubs, link concentration and imbalance between forward and reverse links each appeared more often than expected by chance. Because these features did not have the same stability, we report them separately rather than call the whole network stable or unstable (Supplementary Tables S6a–c).

We use Δz rather than mediation analysis or a correlation adjusted for other variables because cell-composition differences are strongly reflected in blood methylation (module-level correlations, $|r| = 0.65\text{--}0.84$ in our data). If not accounted for, this shared cell-composition signal could make two modules appear related even when the relationship is not specific to those modules.

To reduce the chance that a dependency is an artifact of the analysis, every Δz claim had to pass four checks (Fig. 1c). First, we corrected for testing many pairs within each pre-defined test set (Benjamini–Hochberg, BH). Second, we compared the result with gene sets of the same size and promoter structure but with gene directions assigned at random. Third, we checked that the hallmark score used to split

people into higher and lower groups was not simply tracking age; this used the part of the score unrelated to age ($R_{\perp \text{age}}$). Fourth, we compared the result with randomly drawn gene sets to test whether the finding was specific to the chosen modules. We grade evidence in two tiers: the stronger tier requires the direction of a Δz relationship to agree in an independent cohort. Dose-shape findings—whether the relationship changes steadily across low, middle and high groups—are treated as single-cohort evidence unless independently tested.

Analyses used two cohorts. The discovery cohort (A) is GSE40279 whole-blood 450K ^[9], $n = 656$, ages 19–101. Cohort B (replication) is GSE210255 EPIC ^[10], $n = 1,394$, African-ancestry family structures, ages 20.5–86.7 ($n > 65 = 330$), with gene-level betas recomputed for this study ($19,242 \text{ genes} \times 1,394 \text{ samples}$).

How hallmark scores relate to organ–age patterns

To test whether dependencies cluster around a few hallmark modules, we compared the observed pattern with chance patterns that allow highly connected modules but do not assign modules to layers. An even spread of links would be the wrong comparison because real networks can contain hubs. We tested all hallmark-to-organ pairs ($14 \text{ hallmarks} \times 35 \text{ organ modules}$). In cohort A, 22 pairs passed the four artifact checks and a test that randomly shuffled module labels (SI Fig. 1; Table S1). For the pre-defined organ analysis, 21 of these pairs also passed the correction for testing 490 pairs (Benjamini–Hochberg, BH ^[11]; Fig. 2a). All 21 use the study’s disease-anchored organ gene sets, built from DisGeNET ^[12] and related sources; the claims therefore apply to these gene-set definitions, as discussed below.

Figure 2d shows how the comparison works for individual participants, using the main organ analysis specified before testing (Methods). Because genes were selected for their relationship with age, a higher hallmark mIC means a higher score than expected for someone of that age—not simply that the person is younger. We split participants at the midpoint of this age-adjusted score, creating two groups with similar age distributions. In the higher stem-cell-maintenance group, bone-marrow mIC was more negatively associated with age ($r = -0.241$); in the lower group, the association was close to zero ($r = +0.005$). The same pattern appeared for senescence mIC and kidney mIC ($r = -0.281$ versus -0.023). SI Fig. 2 shows four other examples, including cases where the hallmark score was associated with a more or a less negative organ–age relationship.

The liver provides an example of several hallmarks being associated with the same organ’s age pattern, sometimes in opposite directions. These values come from the main organ analysis specified before testing (Methods). Among participants with higher telomere-maintenance mIC, the liver–age correlation

was less negative than among those with lower mIC ($r_{\text{high}} = -0.107$; $r_{\text{low}} = -0.414$; $\Delta z = +0.307$, $q = 0.007$). Autophagic-renewal (-0.135 versus -0.398 ; $\Delta z = +0.263$, $q = 0.013$) and inflammatory-regulation mIC (-0.142 versus -0.400 ; $\Delta z = +0.258$, $q = 0.014$) showed the same pattern. Higher genome-maintenance mIC showed the reverse pattern: the liver-age correlation was more negative (-0.423 versus -0.136 ; $\Delta z = -0.287$, $q = 0.008$). Here, Δz is the difference between the two correlations, $r_{\text{high}} - r_{\text{low}}$; the reported values agree after rounding. The two liver hallmarks with opposite patterns are conceptually related but share very few genes, so they appear to reflect different measured signals rather than duplicate measurements (Supplementary Table S5).

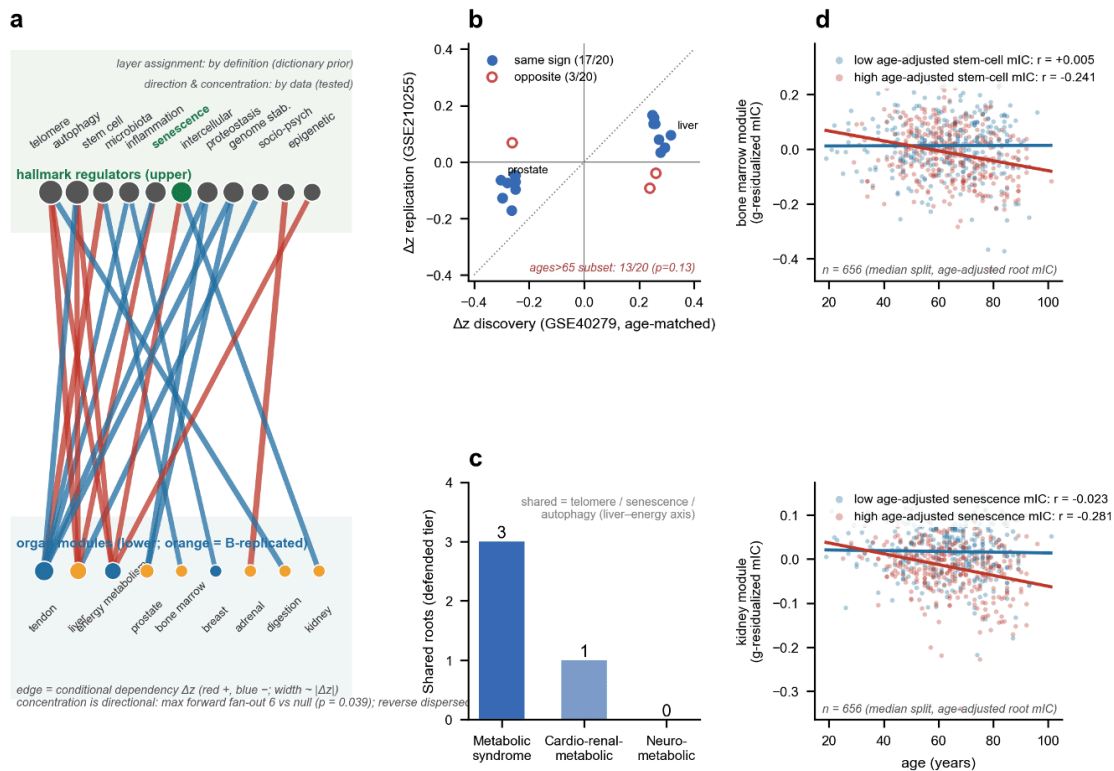


Figure 2. Hallmark scores are associated with organ–age patterns. (a) The network shows the 21 organ pairs that passed the per-pair multiple-testing and age-proxy checks and whose full set passed two gene-set chance tests. Each node’s size shows how many links it has; red and blue edges show positive and negative Δz . Orange organ nodes have at least one link with the same direction in cohort B (10 of 21). Links were more concentrated from hallmarks to organs than in the reverse direction ($\Delta C = +0.18$, permutation $p = 0.030$). The largest hallmark linked to three organs within the 21-pair set; across the broader 490-pair test, the largest forward count was six, more than expected by chance ($p = 0.039$). Reverse links were not concentrated in a few organs ($p = 0.87$; Supplementary Tables S6a, b). (b) In the separate replication list, 17 of 20 pairs kept the same direction in cohort B (one-sided exact binomial $p = 0.0013$, assuming independent pairs with a 50% chance of matching). (c) Few hallmark links were shared across clinical condition groupings. (d) A worked example of the conditional-dependency test: does the association between bone-marrow mIC and age differ according to a person’s stem-cell mIC? Each point is one person in cohort A. We divide participants into higher and lower groups using age-adjusted stem-cell mIC, then compare the bone-marrow mIC–age association between groups. It is more negative in the higher-score group ($r = -0.241$) than in the lower-score group ($r = +0.005$). The same comparison for senescence mIC and kidney shows a similar difference ($r = -0.281$ versus -0.023). This is a cross-sectional example of conditional dependency, not evidence that stem-cell mIC causes bone-marrow aging or a measure of change over time.

To check that the result was not created by dividing participants at the midpoint, we also analyzed hallmark mIC as a continuous score. A regression model tested whether the organ–age relationship changed with the hallmark score, while accounting for sex. This model gave the same direction as Δz for all 21 pairs; 20 of 21 interaction tests had $p < 0.05$. The group-specific correlations and their confidence intervals are available for each pair in the versioned data package.

To see whether the findings held in the independent replication cohort, we checked whether each pair had the same direction there as in discovery (Fig. 2b). Seventeen of 20 pairs agreed in direction. Under a simple chance model in which each pair has an equal chance of agreeing or disagreeing, this is unlikely (one-sided exact binomial $p = 0.0013$); 7 of the 20 pairs also passed the multiple-testing threshold in cohort B ($q < 0.05$). These are two overlapping, not identical, sets: the replication list contains 20 age-matched pairs, while the discovery set contains 22; 13 pairs are in both. Of the 10 organ-line pairs that could be tested in cohort B, all 10 agreed in direction (Table S1). Because cohort B includes related participants from 496 families, we also resampled whole families 1,000 times to account for family structure; the 95% interval for the number agreeing was 13–18. We assess replication by direction, not by requiring the effect sizes to be identical; the size of Δz differs between cohorts (Methods).

The pattern was less consistent among the 330 cohort-B participants older than 65: 13 of 20 pairs agreed in direction ($p = 0.13$). Thus, the replication is seen across the full age range, but has not been established in the older-only group.

The dependencies were not shared equally across all disease groupings (SI Fig. 3). We grouped related clinical conditions using their linked modules; this mapping was assisted by a language model and is marked as such in Methods. Shared hallmark modules were uncommon. In the metabolic-syndrome group (nine linked modules), only three hallmarks passed all four checks: telomere maintenance, senescence regulation and autophagic renewal. All three were linked to the liver–energy axis. The cardiovascular–kidney–metabolic group had one hallmark relationship that also agreed across cohorts (senescence regulation), while the neuro-metabolic group had none.

The analysis narrows in stages: from 332 predefined modules, 108 had enough genes to be scored across the two medical routes. Of these, 94 were function modules (organ, metabolic, immune, intrinsic-capacity and TCM modules); bladder and blood-stasis modules were too small to include. In the main organ analysis, 10 of the 21 pairs could also be tested in cohort B. We then asked how widely each hallmark module was linked to function modules. Stem-cell maintenance, derived from the stem-cell-exhaustion hallmark, had the broadest reach in cohort A: it was linked to 12 modules, compared with a

median of 6 in random-label tests ($p = 0.0001$). Across the 1,316 tested pairs in the four-check analysis, 38 passed the multiple-testing threshold ($q < 0.05$); 67 of the 94 function modules had no significant hallmark link. Under a broader correction across the full hallmark panel, stem-cell maintenance was linked to 13 modules. These counts differ because they come from different sets of tests. In cohort B, 17 of 20 pairs agreed in direction overall, and all 10 testable organ-line pairs agreed.

The counts of 12 and 13 are not contradictory: 12 comes from the stricter four-check analysis, while 13 comes from a broader analysis of the full hallmark panel. The broader count includes links to modules such as bone, bone marrow, brain, heart, lung and tendon, which together connect to 28 clinical phenotypes in our mapping (SI Fig. 3c). That number combines measured module relationships with knowledge-based phenotype mapping; it is a guide to coverage, not a direct test of clinical relationships, and still needs expert review.

We assigned modules to the hallmark or organ layer before testing, rather than choosing the layers to fit the results. We then directly compared links running from hallmarks to organs with links running in the reverse direction. In cohort A, the forward links were more concentrated in a few hallmarks than the reverse links (difference in concentration, $\Delta C = +0.18$; $p = 0.030$). Within the 21 pairs that passed all four checks, stem-cell maintenance was linked to three organ modules. In the broader set of 490 tests, the most-connected hallmark had six links, more than expected in random-label tests ($p = 0.039$; Supplementary Tables S6a, b). Cohort B showed the same direction of difference ($\Delta C = +0.13$), but only three forward links passed the checks, too few for a reliable test.

We also tested two simpler alternatives: a network with equally strong links in both directions, and a broad shift affecting the whole hallmark panel at once. Neither was supported. Although some organ-to-hallmark links were present, no organ consistently acted as a concentrated regulator ($p = 0.87$). When we divided participants by their average score across all hallmarks—or by the main pattern shared across those scores—no individual module remained significant after correcting for all tests (0 of 35 forward; 0 of 14 reverse). The observed pattern is therefore sparse and concentrated in a few hallmark modules, rather than a uniform shift across the whole panel.

One result depended strongly on how the organ gene sets were built (Fig. 2c). With the study's planned disease-anchored gene sets, the modules linked to the most phenotypes—blood vessel (9), brain (9), adaptive immunity (8) and kidney (7)—had almost no detected hallmark links. They appeared as “leaves” in this map. The liver, linked to four phenotypes, had three shared hallmark links in this grouping (four regulators in the detailed results table; Fig. 2a). But when we rebuilt the blood-vessel gene set using a

KEGG-linked definition, four hallmark links appeared: senescence, inflammation, stem cell and nutrient sensing. All four had the same direction in cohort B. These four links did not pass the stricter correction across all gene-set constructions, so this is exploratory evidence that some apparent “leaves” may result from how the organ gene set was defined—not proof that these organs are unregulated. Because blood methylation is strongly affected by cell composition, we treat the leaf pattern as a limitation of this analysis, not as a biological conclusion.

Which organ gene sets carry a measurable signal?

To test whether the organ findings depended on how the organ gene sets were defined, we replaced the disease-anchored sets with organ pathways from Gene Ontology (GO; MSigDB ^[13]). We retested 156 pairs. Across all pairs, directions agreed more often than expected by chance (96/156, or 61.5%; $p = 0.0025$), but none of the 15 pairs that were significant with the disease-anchored sets remained significant with GO sets. Figure 3a compares each of the 21 organ-line pairs across both cohorts and both gene-set definitions (details in Table S2). Six pairs kept the same direction in all four comparisons; stem cell–bone marrow was the strongest ($\Delta z = -0.24, -0.26, -0.15$ and -0.11 ; cohort-B GO $p = 0.035$). Three liver pairs and four energy-metabolism pairs reversed direction in cohort A under the GO definition. We therefore treat the liver-competition and energy-metabolism results as specific to the disease-anchored gene sets. More broadly, these analyses measure dependencies in blood-based signals of disease-related organ aging, not organ identity itself.

We next asked which properties of an organ gene set were associated with detecting a dependency. We scored 151 gene-set definitions using six measures, including how well each matched tissue-specific genes in the Human Protein Atlas (HPA) ^[14], whether its scores were similar in two random halves of the data, and how strongly it tracked the broad blood cell-mixture signal (Fig. 3b–c). None of these three measures positively predicted detection. Instead, the only positive association was with how strongly a module’s score tracked age in both cohorts ($\rho = +0.26$); the strongest negative association was with age information carried by the broad cell-mixture signal ($\rho = -0.38$). Tissue-specific sets most closely matched their named organs and tracked the blood-wide cell-mixture signal least (strict breast set, $|r| = 0.06$), but detected almost no dependencies. Disease-anchored sets had the lowest score consistency across random halves of the data (0.48), yet were the only source with stable detection. In this blood-based analysis, the sets that looked most organ-specific carried the least detectable dependency signal—a reversal of the intuitive expectation. To check whether set size or age-related gene content explained

the disease-anchored advantage, we compared each module with 500 random sets matched for size and 500 matched for age-related gene content. The median proportion of significant findings was 0.15 for the size-matched sets and 0.12 for the age-gene-matched sets, whose directions agreed with the observed ones 43% of the time. In contrast, all seven focal disease-anchored findings remained detectable (7/7). Neither matching approach explained their advantage.

We also tested whether adding protein-interaction neighbors to the disease-anchored organ genes would recover the signal. We added the 20 highest-ranked neighboring genes to each organ set, assigning each added gene to only one organ (Methods). This expanded definition recovered the original pattern rather than replacing it. For the liver, the same four regulators reappeared—autophagy, genomic stability, inflammation and telomere maintenance—and all four had the same direction in cohort B. Eight of the 11 pairs significant under the expanded PPI-based sets were also among the original 21 organ-line pairs, with all eight keeping the same direction. Across 207 gene-set definitions, we counted how often each pair kept the same direction. We use this count to rank pairs, not as a statistical test: the definitions overlap substantially, so together they contain information comparable to only about 1.4 independent comparisons out of 16. We therefore judge support from this ranking together with replication in cohort B using both disease-anchored and PPI-expanded gene sets. Of 39 pair-by-definition combinations, 28 kept their direction in both cohort-B analyses. Stem cell–bone marrow and senescence–kidney ranked highest. A kidney network involving eight hallmarks also kept its direction in seven of eight cohort-B comparisons. It was the largest structure that remained similar across gene-set definitions, although it did not pass the broader correction across the full hallmark panel (Fig. 3d).

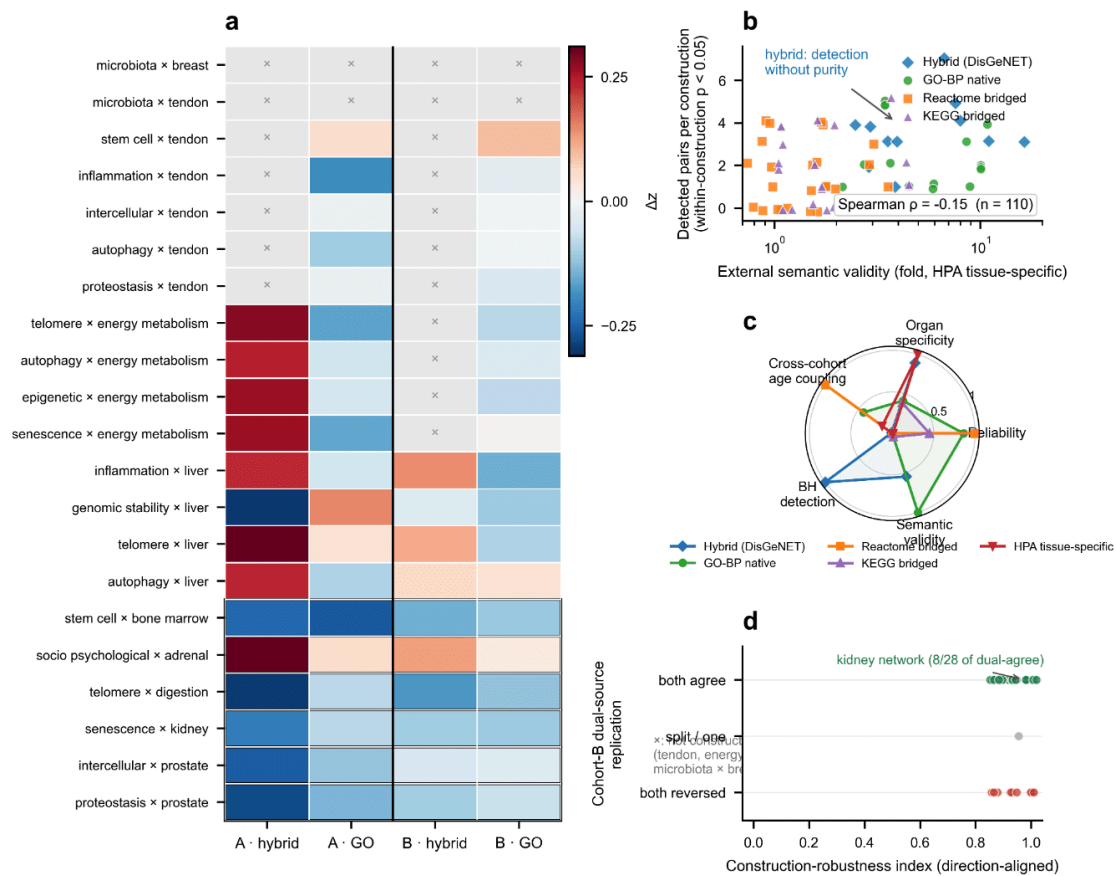


Figure 3. Which organ gene sets reveal a measurable signal? (a) Each of the 21 organ pairs is compared across two gene-set definitions—the disease-anchored set and the GO pathway set—and two cohorts. A black box means the direction agreed in every comparison that could be tested; × marks a comparison that could not be made. (b) The most organ-specific gene sets detected the fewest dependencies, while disease-anchored sets detected the most. (c) No gene-set definition performed best on every measure. (d) For each pair, the plot shows how often its direction agreed across 207 gene-set definitions and whether it replicated using two gene sources in cohort B. Pairs high on both measures have the strongest combined support.

Testing traditional Chinese medicine concepts in the same framework

We asked whether the seven TCM fundamental-substance concepts—qi, blood, essence, yin, yang, spirit and warmth—showed relationships with one another in these scores. We refined the gene lists in five steps to reduce artifacts caused by genes shared across concepts in the STITCH chemical–protein database (Methods; SI Fig. 4). In the final version, genes found in six or more concepts were removed, up to 20 protein-interaction neighbors were added, and each gene was assigned to only one concept. In

cohort B, 10 of the 42 possible pairs passed correction for multiple testing (Fig. 4a; full list in the versioned data package). Essence was linked to five of these ten pairs, more than any other concept.

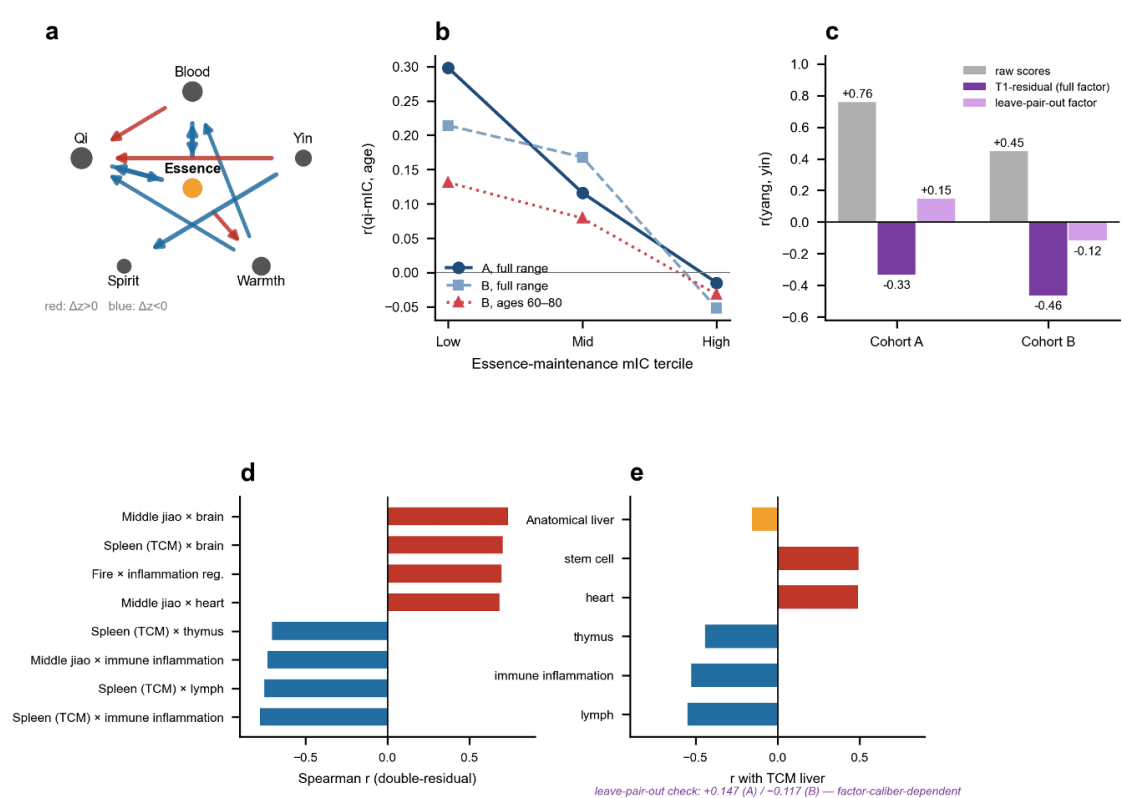


Figure 4. Relationships within TCM concepts and between TCM and modern modules. (a) In cohort B, 10 of 42 pairs among the seven TCM fundamental-substance concepts passed correction; essence was linked to five. (b) For essence–qi, participants were divided into three equal groups by essence mIC, and the qi–age association was calculated in each group. The pattern is shown separately by cohort and age range. (c) Yin and yang scores were positively correlated before adjustment but negatively correlated after removing the pattern shared across the seven concepts. When yin and yang were excluded while calculating that shared pattern, the correlation shifted toward zero (+0.147 in A and –0.117 in B). This result depends on the adjustment and does not show within-person opposition. (d) The strongest relationships across the two systems involve immune modules, including TCM spleen and fire with modern inflammation modules. (e) TCM “liver” was only weakly correlated with anatomical liver ($r = -0.15$), showing that the two labels do not represent the same concept. A1 labels are abbreviated; full names are in Table S10.

Essence–qi was the only concept pair with a dependency in the same direction in both cohorts: $\Delta z = -0.204$ ($p = 0.006$) in cohort A and -0.227 ($p < 0.001$) in cohort B. In cohort A, the association between qi

mIC and age became progressively more negative across groups with low, middle and high essence mIC. The ordering was similar in cohort B across all ages, but weaker when analysis was limited to ages 60–80. The values across low-to-high essence groups were +0.298, +0.116 and –0.015 in cohort A; +0.214, +0.168 and –0.052 in cohort B; and +0.131, +0.079 and –0.032 in the cohort-B age-restricted analysis. This is a measurable association between the two scores, not evidence that the classical idea of essence converting into qi has been confirmed (Fig. 4b).

Yin and yang scores were strongly positively correlated before adjustment ($r = +0.756$ in cohort A), and both declined with age (Spearman correlations –0.62 and –0.63). To separate what they shared with the other TCM scores, we removed the common pattern across the seven concepts. After this adjustment, yin and yang were negatively correlated ($r = -0.332$ in A and –0.463 in B). But when yin and yang were left out while calculating the shared pattern, the negative relationship weakened and changed direction across cohorts (+0.147 in A; –0.117 in B). Thus, the negative correlation depends on how the shared pattern is calculated. These cross-sectional data do not establish that yin and yang restrain one another within a person. At the gene-target level, yang targets were more related to chromatin remodeling and positive RNA regulation, while yin targets were more related to negative regulation and cell-cycle inhibition. These target patterns were reproducible, but the claim of mutual restraint between people was not established.

Two expected relationships were not supported after accounting for shared genes. Qi–blood and blood–spirit did not pass the tests once overlapping genes were taken into account; the qi and blood gene sets shared 47.7% of their genes. Their earlier cross-cohort signals were therefore consistent with the same genes appearing in both sets, rather than evidence for a distinct relationship. The pharmacological target data used to build these modules do not cleanly distinguish qi from blood, which limits what can be concluded.

Where the two systems overlap—and where they differ

To compare the two systems, we correlated every TCM module with every modern module after removing each module family's shared score pattern. This adjustment helps show relationships beyond the pattern common to each family (SI Fig. 5). Of 2,343 cross-system pairs, 74.8% passed correction for multiple testing. The strongest interpretable relationships involved immune modules (Fig. 4d): the TCM spleen score correlated with immune inflammation ($r = -0.78$), lymph (–0.75) and thymus (–0.71). The fire

concept correlated with the modern inflammation hallmark at $r = +0.70$ before adjustment and 0.30 after adjustment ($q = 1 \times 10^{-14}$).

The two systems also differ. After removing each family's shared score pattern, the TCM liver score correlated only weakly with the anatomical liver score ($r = -0.154$). Its strongest correspondences were with lymph (-0.55), immune (-0.53) and stem-cell modules (+0.49); the noisy tendon module was excluded. This shows that the TCM concept "liver" does not simply match the anatomical organ with the same name (Fig. 4e). The TCM fundamental-substance analysis also tested 42 pairs, while the modern organ analysis tested 490 different pairs. Because these are different sets of questions, their counts should not be compared as if they measured the same network density. The two systems overlap in some relationships, but each retains its own pattern.

From clinical problems to candidate interventions

To compare how often interventions changed module scores, we tested each module in each intervention series (Fig. 5b). Across eight intervention datasets (GSE107205, GSE185389, GSE190164, GSE197881, GSE54643, GSE74414, GSE98056 and GSE98876), we analyzed 248 before-and-after participant pairs: 238 were directly paired, and 10 were assembled using stated assumptions. 23.2% of module-series tests changed significantly for T1 modules (13 of 56 tests). For A1 hallmarks, the rate is 14.4% (15 of 104 tests). For each module in each series, we compared scores before and after intervention in the same participants (paired Wilcoxon test), then corrected all 616 tests together using the Benjamini-Hochberg (BH) method. If the correction is limited to the 104 hallmark tests, 13 rather than 15 pass; the two differences are listed in Supplementary Table S7. Figure 5b uses the correction across all 616 tests. Dexamethasone changed five T1 modules in one cohort.

Across all 616 module-by-series tests, hallmark modules had the lowest rate of detected change (14.4%), while organ, metabolic and immune function modules had the highest (25.4%; Fig. 5b). This difference passed a test comparing the two layers (Fisher $p = 0.020$). We also compared the rates within each intervention series, so each comparison used the same intervention and participants. In all seven series where both layers could be compared, the hallmark rate was lower; one additional series had no detected change in either layer. A two-sided sign test excluding that tie gave $p = 0.016$; if the tie is counted as a failure, $p = 0.070$. This is the primary comparison (A1 hallmarks versus A2-A5 functions). Adding the TCM layer gave the same pattern in all eight series ($p = 0.008$; average difference -0.11, bootstrap 95% interval -0.16 to -0.06; Supplementary Table S8). Comparing only hallmarks with TCM modules was a

separate, supplementary test: six of seven non-tied series went in the same direction, but the result was not significant ($p = 0.125$).

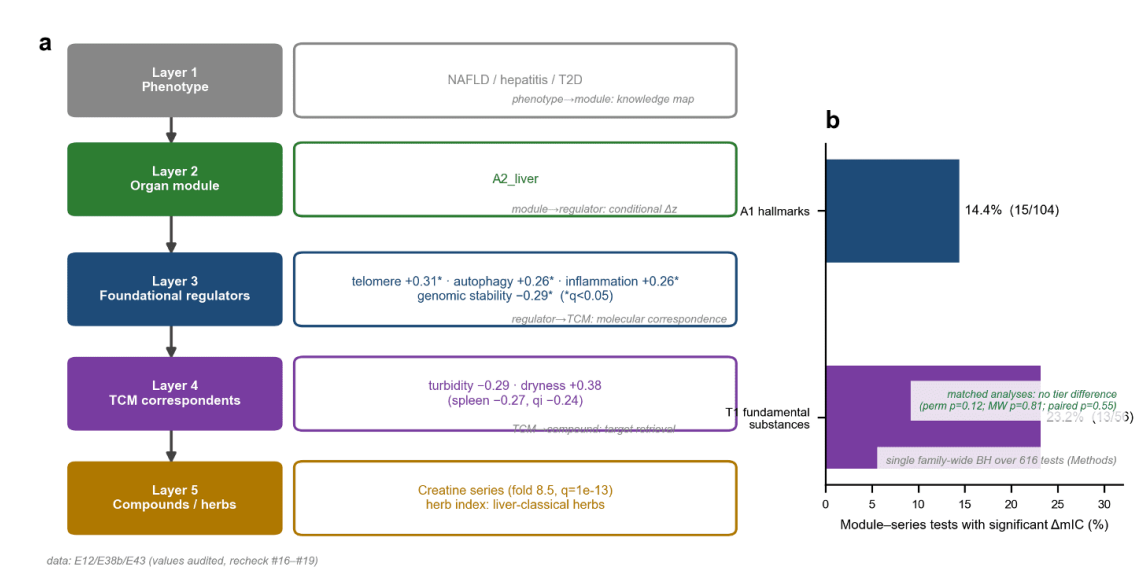


Figure 5. Intervention-related score changes and a route from clinical problems to candidate interventions. (a)

The example traces a liver-related clinical problem through its organ module, associated hallmark modules, TCM concepts, compounds and herbs. This is a way to navigate possible candidates, not evidence that the links are causal. (b) Each test asks whether a module's score changed before versus after an intervention series; ΔmIC is the post-treatment minus pre-treatment score. A positive value means the estimated capability score rose, not necessarily that health improved. Under correction across all tests, significant changes were found in 13/56 TCM tests and 15/104 hallmark tests. A matched TCM-versus-hallmark comparison found no difference (all $p > 0.1$). Hallmarks had fewer changes than organ, metabolic and immune function modules in 7/7 non-tied series ($p = 0.016$), or in all eight when TCM modules were included ($p = 0.008$).

We report these three comparisons as primary, robustness and supplementary analyses, rather than treating them as three interchangeable tests. For that reason, the exact p -values are not adjusted across the three comparisons. In the paired sign tests, series with equal rates are excluded from the main calculation; we also report what happens if a tie is counted as a failure (Supplementary Table S12).

A higher percentage of significant tests does not necessarily mean larger intervention effects; it also depends on how variable the scores are and how much data are available. TCM modules fell between the two modern layers (23.2%). The function layer had the largest median effect size (median $|r| = 0.43$,

compared with 0.35 for hallmarks). Across the eight interventions, the rates of detected change in hallmark and function modules tended to rise and fall together (Spearman $\rho = 0.78$, $p = 0.023$). Every intervention that changed a hallmark module also changed function modules, while four interventions changed function modules without changing hallmarks. This pattern does not show that a change in hallmarks caused the function changes: a test preserving each intervention's number of changes and the number of modules in each layer found no extra nesting beyond chance (10,000 permutations, $p = 0.83$; 8–79 participants per series).

A direct comparison of the TCM fundamental-substance layer (T1) with the hallmark layer (A1) did not find a difference. The module-label permutation test gave $p = 0.12$ (10,000 permutations); effect sizes were also similar (median absolute rank-biserial values 0.30 versus 0.35; Mann–Whitney $p = 0.81$), as were the within-series medians across eight series ($p = 0.55$). This does not conflict with the comparison of hallmark modules against all function layers, which was lower in every series after adding TCM ($p = 0.008$). Here, “accessibility” means only that a module's score changed in the eight intervention types studied; it is not an inherent property of the module.

Figure 5a illustrates how the map can be used to move from a clinical problem toward candidate interventions: phenotype → organ module → hallmark regulators → related TCM concepts → compounds and herbs. For the liver example, telomere maintenance, autophagy and inflammatory regulation were associated with a less negative liver–age relationship, while genome maintenance was associated with a more negative one. The related TCM scores were turbidity (−0.29) and dryness (+0.38); the compound index highlighted creatine-related molecules (enrichment fold 8.5), and the herb index returned herbs traditionally associated with the liver. Figure SI 6 combines the separate analyses into four layers—hallmark regulators, organ functions, TCM concepts and candidate interventions—and labels edges by their evidence type. The compound index uses PubChem Compound IDs and contains 3,125 unique molecules, 2,975 module–compound pairs and 1,154 pairs passing multiple-testing correction (SI Fig. 7). Because both the compound targets and TCM gene modules draw on the STITCH database ^[15], this index is for finding candidates, not independent confirmation that they work.

What the dependency map adds to molecular networks

To see which modules were more likely to change in the intervention series, we compared four module features with the change rates across the 616 module-by-series tests (Fig. 6). Of these features, only a module's layer was informative. The average number of protein-interaction links per module gene did

not predict change (Spearman $\rho = +0.04$, permutation $p = 0.71$). Hallmark modules changed less often than other modules ($\rho = -0.32$, $p = 0.004$; $q = 0.016$ after correcting the four feature tests; Mann–Whitney $p = 0.005$). This difference remained after accounting for gene-set size (adjusted rank correlation = -0.27). A hallmark’s number of linked function modules (fan-out) did not predict how often it changed ($\rho = -0.08$, $p = 0.61$; within A1, $\rho = -0.22$; within A2, $\rho = +0.10$; both non-significant). Nor did the strength of a module’s age association in cohort B ($\rho = +0.04$, $p = 0.73$). Among modules with strong age associations, the relationship was still not significant ($\rho = +0.33$, $p = 0.19$), even though its direction agreed in all 17 cases.

Some modules were linked to many others but rarely changed in the interventions. These included stem-cell maintenance, telomere maintenance, proteostasis maintenance, and intercellular coordination. In contrast, inflammatory regulation had few links but changed readily (Fig. 6d). In these data, molecular-network features did not predict how often a module changed, but its layer did. Thus, being linked to many functions (“influence”) and being changed by the interventions tested (“accessibility”) were different properties; target selection may need to consider both.

We also tested whether interventions changed a hallmark module and its linked function modules in a coordinated way along the specific dependency links (Fig. 6b). If these links carried intervention effects, changes in hallmark and function scores should be more closely related for the 21 tested dependency pairs than for the other 469 hallmark–function pairs. The main comparison was in the expected direction but not statistically clear: the association was 0.39 for dependency pairs and 0.27 for the other pairs (permutation $p = 0.19$; 112 intervention–pair observations, with a wide interval). Other ways of defining coordinated change also showed no difference: direction-based association was 0.24 versus 0.30 ($p = 0.52$), same-direction changes occurred in 67.0% versus 66.4% of pairs, and binary significant-change counts were 8/13 versus 310/497 (Fisher $p = 1.0$). We therefore do not find evidence that intervention effects travel along the specific dependency links. The observation that function modules changed more often overall could be explained by there being more function modules (chance comparison $p = 0.83$). These dependency links should be read as clues to age-related patterns, not as routes along which treatment effects are known to pass.

These intervention results have an important limit: they come from only eight series, so a single series can materially change the comparison. Any claim that effects travel between layers remains a hypothesis, not an established result.

To see which measurements carried across cohorts, we tested whether discovery-cohort measures were associated with the size of each pair's dependency in cohort B (Fig. 6e). Of the 490 hallmark–organ pairs tested in cohort A, 429 could also be tested in cohort B. We compared three measures of molecular proximity—the number of direct PPI links, overlap between the modules' PPI neighbors, and their average network-degree ratio—with four measures from the layered map. None of the PPI measures was associated with cohort-B $|\Delta z|$ ($\rho = +0.01$ to -0.04 ; permutation $p = 0.34$ – 0.88). Three layered-map measures were associated: discovery $|\Delta z|$ ($\rho = +0.153$, $p = 0.005$), the strength of its statistical evidence ($-\log q$; $\rho = +0.154$, $p = 0.005$), and the number of hallmark regulators linked to an organ ($\rho = +0.186$, $p = 0.0002$). The number of organs linked to a hallmark was not associated with transfer ($p = 0.56$).

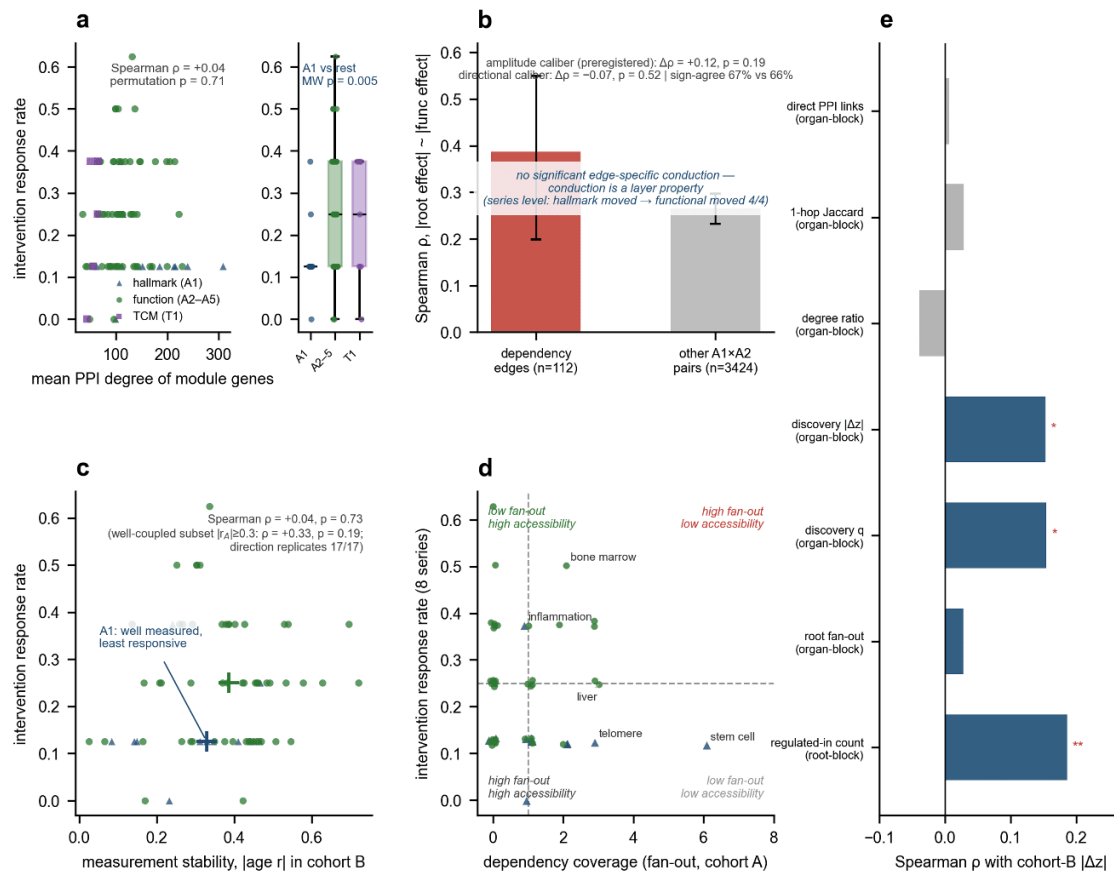


Figure 6. Which module features relate to intervention response or cross-cohort transfer? A module's number of links and its likelihood of changing after intervention are different properties. (a) The average number of PPI links per gene did not predict response ($\rho = +0.04$, permutation $p = 0.71$). Layer did: hallmark modules changed less often than other layers ($\rho = -0.32$, $p = 0.004$; $q = 0.016$ after correcting the four feature tests; Mann–Whitney $p = 0.005$). (b) We found no clear evidence that intervention changes traveled along the specific dependency links. The association between hallmark and function effect sizes was 0.39 for the 21 dependency pairs (112 observations) and 0.27 for the other 469 pairs (3,424 observations); the difference was not significant (permutation $p = 0.19$). Direction-based values were also inconclusive (-0.07 , $p = 0.52$). Changes had the same direction in 67.0% versus 66.4% of pairs; binary counts were 8/13 versus 310/497 (Fisher $p = 1.0$). Every intervention that changed a hallmark also changed functions (4/4), but this was no more frequent than expected from the larger number of function modules ($p = 0.83$). (c) A module's age correlation in cohort B did not predict its intervention response ($\rho = +0.04$, $p = 0.73$); crosses show layer medians. Hallmark modules were strongly age-associated but least responsive. (d) A hallmark's number of linked modules did not predict how often it changed ($\rho = -0.08$, $p = 0.61$). Stem-cell and telomere maintenance had many links but low response. Triangles are hallmarks, circles are organ modules, and squares are TCM modules. (e) Among 429 pairs testable in cohort B, PPI proximity did not predict dependency size ($p = 0.34$ – 0.88), while discovery $|\Delta z|$

and q did ($p = 0.005$). The number of hallmark regulators linked to an organ also transferred when the tested pair was excluded ($p = 0.0023$); the total number of organs linked to a hallmark did not. After accounting for PPI proximity, 102% of the $|\Delta z|$ association remained; after accounting for $|\Delta z|$, PPI proximity was near zero (-0.005).

We next tested whether each measure remained associated with cohort-B dependency after accounting for the other measures. After adjusting for PPI proximity, the discovery $|\Delta z|$ association was essentially unchanged (102% of its original size). By contrast, after adjusting PPI proximity for $|\Delta z|$, its association was close to zero ($\rho = -0.005$). This suggests that, among the measures tested, dependency carried the cross-cohort association; the near-zero value means no association was detected, not that independence was proven. The association also remained after accounting for age association, score consistency across random data halves, gene-set size, and probe coverage for both modules (adjusted $\rho = +0.18$, $p = 0.002$). The count of regulators linked to an organ includes the pair being tested, so we recalculated it without that pair. The count remained associated with cohort-B dependency ($\rho = +0.148$, $p = 0.0023$), but did not add a reliable association beyond discovery $|\Delta z|$ ($+0.07$, $p = 0.139$; Supplementary Tables S9, S11).

We also tested whether the pair-level information could predict dependencies for an organ left out of model fitting. In this held-out-organ test, the logistic model did not show a reliable advantage (AUROC 0.52 versus 0.45; the cluster-based confidence interval included no difference). Thus, the information transferred between cohorts for known pairs, but did not reliably generalize to a new organ.

We also tested six common measures of where a gene sits in a molecular network: betweenness, closeness, PageRank ^[16], k-core, the fraction of a module's genes in the network core, and degree. None predicted the outcomes in either task after correcting across the tests (all permutation $p > 0.05$); positive-control features did pass ($p = 0.004$ and 0.006), showing the tests could detect a signal. Grouping genes into network communities with the Louvain method ^[17] also did not reveal a relationship to the dependency map. Pairs passing the four checks were no more likely than expected by chance to fall in the same community ($p = 0.87$), and shared community did not predict dependency size in either cohort. These findings apply to the network measures and tasks tested here, not to all possible uses of interactome data.

Two other prediction tasks did not work. Module features alone could not predict which modules would change under a new intervention; results were below chance because interventions differed in their overall tendency to change many or few modules. Within a single intervention, layer features ranked

modules in the expected direction only weakly (mean $\rho = +0.08$; positive in 6 of 8 series) and not significantly across eight series ($p = 0.31$). Thus, the useful information was the broad layer a module belonged to, not its rank among modules within that layer. This pattern was similar for the three molecular interventions and five systemic interventions: layer prediction pointed in the same direction in both groups (Fisher $p = 0.084$ and 0.118), while PPI proximity was not predictive in either ($p = 0.29$ and 0.84).

To define where the layered map helps—and where it does not—we compared it with a published drug-repositioning benchmark. We first reproduced the benchmark's PPI-based rankings across 23 drug or disease categories (ATC labels) and 1,916 compounds; every area-under-the-curve score was within 0.01 of the published result. On these tasks, PPI proximity performed well (mean AUROC = 0.825; hypertension = 0.869; type 2 diabetes = 0.920). The layered dependency score could be calculated for only five categories, because the other 18 were drug classes rather than diseases with a defined disease-gene set. On those five tasks, the layered score was weaker (mean AUROC = 0.594); three tasks passed multiple-testing correction—hypertension (0.595), depression (0.672) and atherosclerosis (0.621; $q \leq 0.013$)—while two did not. Adding the layered score to the layer-level averages improved performance by only 0.003, and ranking the top 20 candidates by this score recovered no more than 10% of known hits. The two scores were only weakly related (Spearman $\rho = -0.18$ to $+0.06$). In these tests, molecular proximity and the layered score contributed to different tasks rather than one replacing the other.

One candidate-list test showed a possible use for layer labels, but the result has an important limitation. We filtered PPI-ranked anti-aging compounds to favor candidates linked more strongly to function modules than to hallmark modules. In the top 50, the list became more enriched for known aging-related drugs (3.19-fold to 8.87-fold; $q = 0.033$); three of 18 known hits were aspirin, dasatinib and resveratrol. The result remained when each hit was removed in turn (minimum $p = 0.037$). The filter used PPI links and pre-assigned layer labels—not the learned dependency relationships. At the same list size, the unfiltered top 18 contained one known hit. The gain exceeded label-shuffling ($p = 0.014$) and same-size random lists ($p = 0.039$), but not random lists matched for the number of drug targets ($p = 0.18$). Filtered candidates had more DrugBank targets (median 18 versus 11), which may explain the apparent gain. We therefore interpret this as a possible way to shorten a candidate list, not evidence that dependency edges improve screening. A separate nutrient-layer signal ($p = 3 \times 10^{-4}$) was excluded as an artifact: six of ten sampled positives had helped define the nutrient modules themselves (Supplementary Tables S3, S4).

Discussion

We asked whether a small number of shared aging processes are associated with age-related patterns across specific organ functions. RootMap adds this layer of information to molecular networks: not only which molecules connect, but which foundational capability scores are associated with differences in organ–age patterns. We found relationships in both the modern and TCM module sets, but the two systems overlapped only partly, and observed intervention responses did not establish cause and effect. We therefore report evidence in two tiers: whether a dependency keeps its direction in a second cohort, and whether a dose pattern appears within a cohort. A further limit is how each module’s gene set is built. Across 207 gene-set definitions from seven sources, GO pathway definitions lost the organ signal, while PPI-expanded versions of the same disease-anchored gene sets recovered it. For this reason, each organ claim must be read in light of its gene-set definition. The strongest claims are those that agree across gene-set definitions, cohorts and independent gene sources; the others are explicitly limited to the construction in which they were found.

The scoring method and the individual relationships did not have the same stability. When we kept the scoring rules fixed, module-level age associations kept their direction in the independent cohort. When we reselected genes and recalculated scores, only 7 of 14 testable dependencies kept their direction; repeating the search did not reliably recover the original set of links. An apparent increase in how concentrated the links were came from having fewer pairs to test, not from more links to one regulator. These results argue against a single label such as “stable network.” Future versions should report separately whether scores, individual links, main hubs, link concentration and forward-versus-reverse direction are reproducible.

For a medical world model, RootMap provides a shared set of module scores and measured relationships among them. It does not yet model how those scores change over time or predict a person’s response to treatment; those are separate tasks pursued by population-scale health models ^[4]. Its intended contribution is to suggest candidate points for further study. The results also separate two questions that matter when choosing targets: which modules are linked to many functions, and which modules changed in the intervention datasets. These were not the same modules, and ordinary PPI network centrality did not predict the observed changes. In the drug-screening test, layer labels offered an additional way to organize PPI-ranked candidates, but the gain was not distinguishable from lists matched for drug-target count. In the cross-cohort test, dependency measures—not the tested PPI-proximity measures—tracked

replication, but the association was modest (ρ about 0.15). This ranks pairs; it does not sharply separate them or prove that PPI measures are unrelated.

The drug-repositioning benchmark sets a boundary on this interpretation. For the tested ATC-labeled drug and disease categories, PPI proximity performed well; the layered score added little beyond the layer-level averages. We therefore do not claim that RootMap replaces network medicine. The two approaches answer different questions in these tests: PPI proximity helps rank candidates in some drug-repositioning tasks, while RootMap adds information about how capability states relate to age patterns in organs.

Two findings need careful interpretation. First, the hallmark–organ relationships were mostly organ-specific rather than shared across disease groupings. In the disease-anchored analysis, blood vessel, brain, adaptive immunity and kidney had many linked phenotypes but few detected hallmark relationships. A KEGG-based version of the blood-vessel gene set produced four candidate links, but these did not pass correction across all gene-set definitions. This shows that an apparent lack of links may depend on how a gene set is built; it does not prove that these organs are unregulated. Nor do the results show that changing a hallmark would improve multiple organs. Second, the measurable signal depended on the gene sets used. The most tissue-specific gene sets detected little, while disease-anchored sets and their PPI neighbors carried more of the signal. In blood methylation, the data may reflect disease-related changes in an organ more than the organ's identity. This limits what the organ-level results can be taken to represent.

The adapted intrinsic-capacity panel—cognition, emotion, movement, sleep, sensory function and blood-cell formation—showed 79 strong relationships with the other module set. It was inspired by, but is not the same as, the World Health Organization's Intrinsic Capacity framework [\[18\]](#). TCM concepts could also be tested in the same data, but this does not confirm traditional theory. The yin–yang relationship depended on how shared score variation was removed, while the essence–qi association kept its direction in both cohorts. These are measurable patterns in the module scores, not proof of classical mechanisms.

Organ-specific aging clocks provide an important comparison. Proteomic clocks built across populations show that organs can age at different rates and that these differences predict disease [\[19\]](#). They do not, however, show which hallmark states are associated with each organ's age pattern—the question addressed by RootMap. Separately, estimates suggest that changing blood-cell composition explains up

to 39% of the accuracy of blood-based aging clocks [20]. This supports our decision to account for cell-composition effects when analyzing blood methylation.

Aging-hallmark frameworks organize processes that may contribute to aging [2][3], but often leave relationships among those processes as hypotheses. RootMap makes one such relationship measurable by asking how many organ modules are linked to each hallmark. The broadest hallmark differed by cohort: stem-cell maintenance in cohort A and senescence regulation in cohort B; telomere maintenance was not among the broadest in either. This is consistent with network studies showing that hallmark genes form connected groups in molecular interaction maps [1]. Our construction tests add that, in blood methylation, the signal was more detectable in disease-anchored genes and their PPI neighbors than in pathway-ontology gene sets. In TCM, network pharmacology has often studied one formula for one disease [21]. Here we place TCM's fundamental substances and modern hallmarks in one scoring framework and apply the same dependency test. Findings such as essence's hub position describe the organization of these scores; they do not establish that a particular formula treats a disease.

Population-scale health models that simulate future health trajectories are another related field [4]. They are not competitors to RootMap: they predict how health states may change, while RootMap tests how measured capability scores are related to organ-age patterns. Such models could eventually use these dependencies as information when estimating transitions or choosing interventions. In our transfer analysis, the tested PPI-proximity measures did not predict the cross-cohort dependency association, whereas measured dependency statistics did. This supports studying the two kinds of information together, but does not show that connectivity is useless or that a full transition model can be built from RootMap alone.

Our study has several limits.

1. The analyses are observational. A “dependency” here means that an organ score's association with age differs across levels of another module; it does not show that one module causes change in another. We use “conditionally regulate” only as shorthand for this measured difference.
2. Agreement between cohorts in the direction of a dependency is not the same as having mIC scores on the same numeric scale. The fixed reference scale does not yet calibrate absolute mIC values across cohorts (Tables S13, S14). Out-of-range scores track shifts in the average PAI score, but platform and population differences are confounded. We therefore compare directions, ranks and

paired changes across cohorts; direct comparison of absolute mIC values requires a separate calibration study.

3. The modules are measured in blood, where cell-composition differences strongly affect methylation. The apparent lack of links for some high-phenotype-burden organs may partly reflect the difference between blood and the organ tissue itself.
4. Organ findings depend on how the organ gene sets are defined. The liver-competition and energy-metabolism patterns reversed with pathway-based gene sets, so we limit those findings to the disease-anchored sets. Comparing many gene-set definitions can rank pairs, but cannot provide an independent statistical test because the definitions reuse data (estimated independent information is about 1.4 comparisons out of 16). We therefore rely on separate cohort and gene-source replication. Some relationships that agreed across gene-set definitions—including extracellular matrix, nutrient sensing and socio-psychological roots—changed direction in cohort B and need testing in a third cohort.
5. TCM modules were built from pharmacological target lists, not clinical definitions of syndromes. The qi–blood relationship failed after shared genes were removed, showing that the source data do not clearly separate those concepts. The pathological-pattern modules (T3) remain scores for named concepts; their meaning as measurable capabilities has not yet been established.
6. The links between clinical phenotypes and modules were generated with a language model and still need expert review.
7. Directional agreement was weaker among participants older than 65 (13/20 pairs). Dose patterns also differed across cohorts and age ranges.
8. The compound index and the TCM modules draw on the same STITCH data source. The index can suggest compounds to investigate, but does not independently validate them.

These limits point to the next steps. Adding protein, imaging and other clinical measurements could test whether the same dependencies appear in different tissues, populations and time points, and whether they persist after an intervention. The next task is to move from a map of relationships to a model of how states change over time. Such a model could use the dependencies measured here as candidate relationships to test—not as established routes through which treatments act.

RootMap is an initial map, not a complete dependency network. Only 21–22 relationships passed the current checks, in two blood-methylation cohorts and one measurement type. Proteomic organ clocks, imaging, and longitudinal intervention studies could test or add relationships where blood data are

insufficient, using the same dependency criteria. New evidence may support, revise, or rule out individual links, so the map is designed to remain open to change.

For longevity medicine, the goal is not only to predict age-related change but to choose what to measure and which interventions to test as people seek to preserve several functions over time. RootMap contributes an early, evidence-graded map for that research: it links measurable maintenance states to organ-aging patterns and distinguishes candidate breadth from observed intervention responsiveness. It may help prioritize hypotheses for longitudinal or N-of-1 studies, but it does not predict an individual's response, identify a proven intervention target, or establish clinical benefit. This is consistent with proposals that health prediction should support intervention choice [22] and that N-of-1 decisions in longevity medicine need explicit evidence standards [23]. Calibrated mIC scores and replicated dependencies could eventually inform those decisions; at present, absolute scores still need cross-platform and cross-population calibration.

Methods

A reader's guide to the terminology

This paper distinguishes three levels. A phenotype is a clinical condition or problem. A specific function is represented by an organ, metabolic, immune, or intrinsic-capacity module. A foundational capability is a process—such as maintenance, repair, or renewal—that may support those functions. In this study, aging-hallmark-derived modules represent the foundational level.

Each module is a predefined list of genes. We score it in each blood sample using methylation in the gene promoters (the regions that help regulate gene activity) for genes whose methylation rises or falls with age. To test a dependency, we compare the association between a function score and age in people with higher versus lower age-adjusted scores for a hallmark module. The difference between those age associations is Δz . “Regulator” simply names the hallmark module being used to split the comparison; it does not mean that the module has been shown to cause a change in the other module.

We use “analysis setup” for the exact set of pairs tested and the statistical choices used for a result. The paper sometimes calls this a “caliber”; two analyses can give different counts because they test different pairs or use different corrections. A “gene-set definition” (also called a construction) is one way of choosing the genes for a module, such as a disease-anchored set, a GO pathway, or a tissue-specific set. Comparing definitions tests whether a result depends on how the genes were chosen. We call a result

replicated when its direction is the same in an independent cohort. Patterns across low, middle, and high groups are called dose shapes and are treated as single-cohort results unless tested independently. Intervention analyses compare module scores before and after treatment; they do not show that an intervention reversed aging or disease.

Study cohorts and sample processing

We analyzed two whole-blood methylation cohorts. Discovery cohort A was GSE40279, measured on the Illumina 450K array (656 participants, ages 19–101). Replication cohort B was GSE210255, measured on the EPIC array (1,394 participants from African-ancestry families). For cohort B, we recalculated methylation scores for 19,242 genes across all 1,394 samples by averaging methylation at probes near each gene's promoter. Promoter probes were defined as UCSC RefGene Group TSS1500/TSS200 using the Illumina 450K v1-2 manifest. Cohort B was limited to promoter probes shared by both platforms so the cohorts could be compared on the same measurements.

The counts below come from different analysis steps and test sets, so they should not be read as one single sample count: the 332-module dictionary yielded 108 modules after the minimum-gene filter; the main analysis included 94 function modules (A2–A5 and T families); the organ analysis tested 21 pairs, of which 10 could be tested in cohort B; the broader four-check discovery analysis identified 22 pairs; and the separate calibration audit covered 105 cohort-B modules (Table S13). The fan-out counts of 12 and 13 also come from different test sets. Intervention analyses used 248 before-and-after sample pairs from eight series, taken from the existing `delta_wide_248` dataset without reprocessing.

Gene modules and how we score them

The dictionary was fixed on 2026-09-05 and contains 332 modules. It operationalizes the panoramic framework described in the Results: 14 aging-hallmark modules (A1) form the proposed foundation; the three functional pillars are organ (A2, 36 modules), metabolism (A3, 10), and immunity (A4, 6). Together, these four groups comprise the 66-module modern system. Six additional intrinsic-capacity modules (A5) extend coverage to broader whole-person function. A parallel TCM system contains 38 concepts. The remaining 222 modules are for nutrients (NUT, 84; NUTX, 38) and food-herb items (YFY, 100), used for phenotype mapping and candidate retrieval. In this study, Capomics refers to deriving module-level intrinsic-capability (mIC) scores from blood DNA methylation. The promoter-aggregated index (PAI) compares methylation in genes whose promoter methylation rises with age with methylation in genes

whose promoter methylation falls with age. We use Illumina beta values directly in both cohorts, without converting them to M-values. The age-related gene lists were selected in cohort A using an absolute Spearman correlation of at least 0.2 with age, a threshold set before outcome analysis; the same lists were then used in cohort B. Equation 1 gives the exact score for each participant and module:

$$PAI(i, m) = \frac{\bar{\beta}(pos) - \bar{\beta}(neg)}{\sqrt{var(pos)/n_{pos} + var(neg)/n_{neg} + 10^{-12}}} \quad (1)$$

In Equation 1, β is the methylation value for a gene's promoter, and n_{pos} and n_{neg} are the numbers of genes in the two age-related groups. The denominator accounts for how much the genes vary within a sample and how many genes are available: more variation makes the score less precise, while more genes make it more precise. A module is scored only when at least two genes are available in each group. The PAI formula does not use cohort-wide centering or scaling.

No gene values were missing in the analyses reported here, so imputation was never needed (0.000 imputed fraction in both cohorts; module-level audit E99, versioned data package). Cohort B used the same formula and the age-related gene lists fixed in cohort A. The normalization bounds L and U were set on the PAI scores in cohort A, not on the raw β values. For intervention data, the change in mIC was calculated from the change in PAI as $\Delta mIC = -\Delta PAI / (U - L)$. mIC ranges from 0 to 1 only in the reference cohort. In other cohorts, we apply the same linear conversion without cutting off values outside that range. This keeps ranks, correlations, and paired-test results unchanged, but does not ensure that absolute mIC values are calibrated across cohorts. Supplementary Table S13 reports each module's observed range and how many values fall outside 0–1.

How we calculate a dependency (Δz)

To calculate a dependency between a hallmark module (R) and a function module (F), we first divide participants into higher- and lower-R groups using the midpoint of the age-adjusted R scores. We then calculate the correlation between F and age in each group. The difference between these two correlations is the effect size Δz (Equation 2):

$$\Delta z = r_{high} - r_{low} \quad (2)$$

This represents the difference between the two within-group correlations. Inference uses the two-independent-correlations Fisher-z statistic (Equation 3):

$$z = \frac{atanh(r_{high}) - atanh(r_{low})}{\sqrt{1/(n_{high}-3) + 1/(n_{low}-3)}} \quad (3)$$

We report Δz (the difference between correlations) separately from z (the test statistic used to calculate its p-value). P-values are two-sided and corrected for multiple testing within each stated set of pairs using Benjamini–Hochberg (BH). Before splitting participants, we adjust module scores in three ways: remove the pattern shared within each module family; adjust pairs that share genes before testing them; and remove both the shared family pattern and age-related component from the hallmark score used for the split. These steps reduce the chance that a shared family signal, overlapping genes, or age itself explains a result. The data package gives each pair’s high- and low-group correlations and their confidence intervals. As a second check, we tested the hallmark score continuously in a model including age, the hallmark score, their interaction, and sex; all pairs had the same direction as Δz . The cross-fitting analysis that repeats gene selection is described above. For cohort B, uncertainty in the number of pairs with matching directions was estimated by resampling whole families 1,000 times.

Module-level intrinsic capability (mIC). We build on the capability index defined in our steerability framework ($CI = 1 - PAI$) [5]. Each module score is expressed as (Equation 4):

$$mIC = 1 - N(PAI) \quad (4)$$

where N is a linear min–max normalization with parameters L_m and U_m frozen on the discovery cohort ($n = 656$). Within the reference cohort, mIC values fall within [1], and higher values indicate greater estimated capability on this scale; functional and clinical validity are evaluated separately. Cross-cohort and intervention analyses use the same frozen parameters without rescaling. Under this fixed negative-slope affine map, the reported statistics are invariant: correlation signs reverse with unchanged magnitude, and Δz , Δr , and two-sided p-values are unchanged over all 1,316 A1×function pairs. The persisted equivalence audit gives machine-precision evidence: $\max |\Delta z_{mIC} - \Delta z_{PAI}| = 1.3 \times 10^{-14}$, $\max |\Delta p| = 8.9 \times 10^{-15}$, zero pairs with sign-flipped Δr , and high/low group-exchange consistency on all 1,316 pairs. This audit covers the downstream affine transform only; it does not by itself certify that the upstream PAI implementations produce identically calibrated scales across cohorts.

One calibration audit (Supplementary Table S13) addresses that gap and shows it is substantial: under the frozen bounds, 31 of 105 cohort-B modules lie entirely outside [1], 53 exceed 90% out-of-bound, and 8 have no out-of-bound samples; the frozen map does not currently achieve a common numeric calibration of mIC across cohorts. A source audit (Supplementary Table S14) shows that, across modules, out-of-bound rates track PAI mean-level shifts more strongly than direction-gene retention (median 0.97) or SD ratio (median 0.72) in these cohorts, a pattern consistent with mean-level scale displacement, with platform and population remaining confounded, and cohort-mean imputation never triggers (E99), so

the attribution remains open pending a third cohort or technical-reference design. We therefore grade dependency transfer and absolute-state calibration separately. Selected conditional relationships transfer across cohorts, whereas many module-level mIC coordinates require further cross-platform and cross-population calibration. Outside the frozen reference cohort, mIC values are extrapolated coordinates rather than bounded capability percentages; their absolute interpretation awaits cross-platform calibration. We do not truncate or rescale per cohort (truncation destroys information and per-cohort rescaling would mask the difference); every module's observed range and out-of-bound counts are reported in Supplementary Table S13.

For intervention contrasts, paired-test p-values are unchanged with sign-reversed effect sizes ($\Delta\text{mIC} = -\Delta\text{N}$). Age-adjusted residuals are analyzed separately and are not bounded to ^[1] (Supplementary Table S10).

Concentration and direction asymmetry (in-sample): concentration C is the maximum fan-out fraction among significant forward edges. The degree-preserving direction-flip null keeps every significant edge and its endpoint degrees fixed, reassigning only which side originates the conditioning (edge-orientation permutation, 10,000 draws). ΔC , the direct difference, compares forward minus reverse concentration under the same permutation framework.

Organization-stability redetection (matched comparison): concentration C is the maximum fan-out fraction among detected forward edges. Two calibers are kept distinct: the fixed candidate set, and redetection on the same evaluable family (BH within family). Each is compared against an edge-multinomial null matched to its own detected edge count and evaluable label pool; twenty random five-fold resplits give intervals. A fold-version reshuffling control breaks the shared-fold structure: per module and fold, the held-out block is scored by a different fold's direction-gene set, leaving data and gene sets unchanged. Marginal age correlations are not reduced by the control, which localizes the effect to the conditional-dependency structure. An isolation audit of this control: in 245 of 245 module-by-held-out-block combinations, the substitute training fold still contains the evaluated participants (structural under five-fold cross-fitting), and threshold-based gene selection is nonlinear, so individual-participant influence cannot be summarized as a uniform weight. The control therefore alters the training-test isolation as well as the version sharing; it is reported as an auxiliary observation in Supplementary Table S6, and stability conclusions rest on the matched redetection comparison and the twenty resplit intervals, with the reshuffling control treated as auxiliary (above).

Four checks against false findings

Each dependency claim had to pass four checks. First, we corrected for testing many module pairs at once using BH within the stated test set. Second, we compared the result with gene sets of the same size and promoter structure but with gene directions assigned at random. Third, we removed the part of the hallmark score explained by age, so the high/low split would not simply separate younger from older participants. Fourth, we compared the full set of findings with results from randomly chosen gene sets of the same size. A pair is called “defended” when it passes the first and third checks individually and its full claim set also passes the second and fourth checks. For the two set-level checks, we ran 1,000 label permutations while keeping the number of positive- and negative-direction genes fixed; both gave empirical $p < 0.001$. During internal review, we added a fifth check for organ findings: test whether they persist when the organ gene sets are built in alternative ways.

Alternative gene sets and how we compared them

The main organ analysis used the disease-anchored gene sets in the 332-module dictionary, built from DisGeNET and related sources. We compared them with two alternatives. The first used organ pathways from Gene Ontology Biological Process (GO-BP), selected by matching organ names. The second started with those GO sets and added genes from Reactome and KEGG Medicus ^{[24][25]} when enrichment tests supported the match ($q < 0.05$); for these sets, we tested all genes and two size limits (top 200 or top 500 genes by frequency across pathways). We also used tissue-specific gene sets from the Human Protein Atlas (HPA) version 25.1, defining strict sets as at least five times more expressed in one organ and enhanced sets as at least twice as highly expressed as in the next-highest organ.

We compared gene-set definitions on six broad dimensions: (1) coverage—how many genes and modules were included; (2) measurement—whether scores were similar across 50 random half-splits, how strongly they tracked the broad blood cell-mixture signal ($|r(\text{PAI}, g)|$), and how much age information remained after accounting for that signal; (3) generalization—whether age associations kept the same direction across cohorts; (4) detection—whether dependencies passed statistical correction; (5) similarity to the main disease-anchored gene sets; and (6) match to tissue-specific genes in the Human Protein Atlas. Detection was corrected within each gene-set definition and across all 1,677 construction tests. The main organ analysis corrected 490 pairs; construction-specific analyses corrected 13; exploratory construction searches corrected all 1,677. A q value depends on how many tests were included, so the

same-looking q columns from different analyses are not directly comparable; some test sets were up to 40 times larger than others.

We also created PPI-expanded gene sets using the STRING interaction network [26]. For each organ, we ranked directly connected genes by how many links they had to the original organ genes, then added the top 20. A gene was assigned to only one organ—the one with the most supporting links; ties were dropped. We used three starting gene sets: the full disease-anchored set, genes supported by at least two sources, and the union of the disease-anchored and strict HPA sets. Across all 207 gene-set definitions, we calculated how often each pair kept the same direction. Because many definitions use overlapping data (median correlation between definitions, $|r| = 0.67$), this fraction is only used to rank pairs, not as a statistical test; it contains about 1.4 independent comparisons' worth of information out of 16. Evidence is therefore judged using independent replication in cohort B across both the disease-anchored and PPI-expanded gene sets.

How the TCM gene sets were refined

The first TCM gene lists (v2) came from STITCH herb-compound target data [15], capped at the top 200 targets per concept. This caused some highly connected genes to appear in as many as 27 of 38 concepts. We refined the lists in five steps. Version 3 added the top 20 PPI-linked genes; version 4 removed shared hub genes and reclassified the T2 group. In version 5, assigning every gene to a single concept failed: 53% of genes were tied across concepts, and no genes were uniquely assigned to blood, a limitation of the source data. Versions 6 and 6+ removed genes appearing in six or more concepts and added PPI neighbors with exclusive assignment. The final sets had low overlap (Jaccard similarity 0.028–0.029; SI Fig. 4). We use v6+ (frequency cutoff $K = 6$, top 20 PPI neighbors, exclusive assignment), selected from a full grid of possible settings. The qi-blood result after removing shared genes is reported as a failure. Pathological-pattern concepts (T3; including fire, toxin, dryness and deficiency) keep their original names. Their scores are not automatically interpreted as modern capabilities—for example, a high “toxin” score does not mean stronger detoxification.

Comparing scores across the TCM and modern modules

After filtering out gene sets that were too small, we compared 33 TCM modules with 71 modern modules (A1 14, A2 35, A3 10, A4 6 and A5 6), for 2,343 pairs. We calculated Spearman correlations after removing the score pattern shared within each module family, then corrected across all pairs using BH. The tendon

module (7 genes) was excluded from headline claims because its score was noisy; it is flagged wherever it appears.

How intervention-related score changes were tested

For each module in each intervention series, we tested whether the score changed from before to after treatment using a paired Wilcoxon test [27], which compares measurements from the same participants. The main correction for multiple testing was across all 616 module-by-series tests; we also report a correction limited to the A1 hallmark tests where it changed the counts (two tests). To compare layers within a series, we calculated the difference between their rates of significant change. Exact two-sided sign tests were used. A series with equal rates was excluded from the main test; we also report an alternative that counts a tie as a failure. The tie identities and both calculations are in Supplementary Table S12. To test whether changes were more often nested across layers than expected, we shuffled significance labels within each intervention while preserving its total number of changes and the number of modules in each layer. A matched comparison of T1 and A1 found no difference in counts, effect-size distributions or within-series medians (all $p > 0.1$).

How we compared module links with intervention response

For each module, we defined its intervention response rate as the share of intervention series in which its score changed significantly after correcting across all tests. We tested whether that rate was associated with four features: the average number of PPI links per gene; the number of other modules it was linked to (fan-out, using cohort-A relationships); how strongly its score was correlated with age in cohort B; and its layer. We used Spearman correlations, 10,000 permutations to assess chance, bootstrap 95% confidence intervals, and an adjusted rank correlation to account for gene-set size. The four feature tests were corrected together using BH. For the separate test of whether effects followed specific dependency links, we compared how closely hallmark and function effect sizes moved together on the 21 defended pairs versus other A1×A2 pairs. The pre-specified main measure was the size of the paired effect, with direction and same-sign measures reported as additional checks. The fan-out and intervention response analyses used different participants.

How we tested what transfers between cohorts

To test which measurements carried over to cohort B, we asked whether features of each pair predicted the size of its dependency there. Of the 490 hallmark–organ pairs in discovery, 429 could also be tested in cohort B. We compared three PPI-proximity measures: the number of direct links between the two gene sets, adjusted for their sizes; the overlap between their first-neighbor genes (Jaccard similarity); and the ratio of their average network degrees. We also tested four map-based measures: discovery $|\Delta z|$, the strength of its statistical evidence ($-\log_{10} q$), how many organs were linked to the hallmark, and how many significant hallmark regulators were linked to the organ. P values came from 10,000 permutations. We shuffled within organs for pair-specific features and within hallmarks for features that were constant for each organ.

For these tests, we shuffled the outcome values within groups of related pairs, breaking the link between each feature and the outcome while preserving the grouping structure. To ask whether PPI proximity and dependency measures each contributed information beyond the other, we used a rank-based partial correlation: we removed the part associated with the other measure, then tested the remaining association using permutations within the same groups. We also adjusted for each module's age association, score consistency across 50 random data splits, gene count and probe coverage. A separate five-fold test left out whole organs in turn and tested whether a logistic model could predict dependencies for an unseen organ. Finally, we tested whether a score summarizing all 14 hallmarks—or their main shared pattern—was associated with each organ score, and also tested the reverse direction. Before this comparison, each hallmark score was standardized, adjusted for the broad shared family pattern and age, and then split at its midpoint. Significance was tested by randomly re-splitting participants 10,000 times. These tests ask whether a panel-wide shift is associated with organ scores, beyond the pattern common to the family.

We also tested two prediction tasks that did not succeed: predicting whether a module would change across a new intervention series, and ranking modules within each intervention. The first was dominated by how strongly each intervention changed modules overall; the second showed only a weak trend across eight series. We compared six other measures of network position—degree, betweenness (estimated from 500 sampled source nodes), closeness, PageRank, k-core number and the fraction of genes in the network core. Each measure was tested as a module average and as a difference between the two modules in a pair, with correction across the full test family. Layer labels and discovery $|\Delta z|$ served as positive controls and showed that the tests could detect a signal. We also grouped genes into network communities using the

Louvain method (resolution 1.0, fixed seed), assigned each module to its most common community, and tested whether related modules shared communities or had similar dependency sizes. Finally, we repeated the layer-versus-PPI comparison separately for three molecular interventions (dexamethasone, omega-3 and flavanol supplementation) and five systemic interventions (Mediterranean diet, vaccination with ART, CPAP, alcohol abstinence and glucose education).

For the external drug-repositioning test, we first reproduced a published benchmark that ranks drugs for 23 ATC-labeled conditions across 1,916 agents. We followed its logistic-ranking setup (L2 regularization, $C = 0.1$, five-fold cross-validation repeated with three random seeds, averaging predictions for held-out cases). Our results matched the archived benchmark to within 0.01 AUROC for every task, with the same number of positive cases. We then calculated a layered score for each drug in disease tasks: the fraction of that disease's significant hallmark regulators that were also linked to modules targeted by the drug. This score could be calculated for only five tasks, because the other labels were drug classes without a disease gene set. We filtered PPI-ranked lists of 50–500 agents to favor those with greater proximity to function modules than to hallmark modules. Enrichment for DrugAge compounds ^[28] was tested with a one-sided Fisher exact test, with correction across eight primary tests. We also checked whether results held when each hit was removed, tested a positive subset not used to define the network (OpenGenes level 1-disjoint subset ^[29]), and compared the filter with same-size random lists, shuffled layer labels and lists matched for drug-target counts (Supplementary Tables S3, S4).

Planned hypotheses and analyses that did not hold

We report the original planned hypotheses (H1–H4: shared variation across layers, unequal forward and reverse links, mediation and intervention transmission) and their 11 tests in the versioned data package. Those tests did not support the original hypotheses, so the present paper focuses on the later analysis of conditional dependencies. Supplementary Table S15 labels each analysis as originally planned, added after the change in direction, or exploratory, and states its evidence level. Six exploratory findings were downgraded because of artifact checks or lack of replication; they are identified as single-cohort findings in the text and supplement.

Linking clinical conditions to modules

We used the language model Claude to link 81 clinical conditions to 112 module anchors, ordered by ICD-11 chapters. These links were generated from medical knowledge and have not yet been reviewed by

experts. Counts based on them are therefore indicators of map coverage, not validated clinical mappings. Conditions that could not be linked (eye, ear, peripheral nerve and polypharmacy) are reported as limits.

Statistics and reporting

For each claim, we report the size of the observed relationship and its q value, which adjusts for the number of tests; we do not rely on a p value alone. Because the data are observational, “conditionally regulate” means only that a module’s association with age differs across levels of another module—it does not imply causation. Results across low, middle and high score groups are labeled as single-cohort unless tested in another cohort. Analyses used Python (pandas, scipy and statsmodels); scripts and analysis logs are archived with their inputs and provenance.

Statements and Declarations

Funding

This research was conducted and funded by DeepoMe Inc. as in-house scientific research.

Potential Competing Interests

The author J.X. is the founder of DeepoMe Inc. (Beijing, China), which develops DNA methylation-based health assessments and related products and services. This research was conducted and funded by DeepoMe Inc. as in-house scientific research. No other financial or non-financial competing interests are declared.

Ethics

This study used only publicly available, de-identified secondary datasets (GEO accessions GSE40279 and GSE210255; Human Protein Atlas v25.1). No new human participants, samples, or personal data were collected by the authors, and no institutional review board approval was therefore required. The intervention-associated datasets listed in Methods are published, de-identified aggregate results from their original studies.

Author contributions

J.X.: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing – original draft preparation, writing – review and editing, visualization, project administration.

Data Availability

The discovery cohort (GSE40279) and replication cohort (GSE210255) are publicly available from the NCBI Gene Expression Omnibus. The Human Protein Atlas consensus tissue transcriptome matrix (v25.1) is publicly available from the Human Protein Atlas. All derived gene-set constructions (207 constructions), module scores, and per-pair statistics underlying the figures, including the figure-level source data for Figs. 1-6 and SI Figs. 1-7, are provided in a versioned data package (Supplementary_Data.zip; 28 files with a checksummed MANIFEST, prepared at submission time) that is available from the corresponding author and will be deposited with a subsequent version of this preprint.

Code Availability

All analysis pipelines (dependency testing, four-line artifact defense, construction-sensitivity analysis, and the six-layer construction scorecard) were implemented in Python 3.11 (numpy/pandas/scipy). The following assets are assembled in the versioned data package (Supplementary_Data.zip; 28 files with a checksummed MANIFEST, script inputs listed in the MANIFEST), available from the corresponding author and to be deposited with a subsequent version of this preprint: the frozen 332-module dictionary and mIC calibration parameters (L, U per module; Table S10), PAI and mIC module scores for the discovery cohort, per-pair dependency statistics for all 1,316 A1×function pairs, the main-caliber per-pair $r_{\text{high}}/r_{\text{low}}$ record behind Figures 2d and SI Fig. 2, the three intervention contrasts with tie rules (Table S12), the persisted mIC affine-equivalence audit, the external calibration audit with its scale-shift and imputation audits (Tables S13, S14; E99), the re-split details behind Tables S6a-c, and the scripts reproducing the main figures and these tables. A complete, versioned archival package with a persistent DOI (access portal at steeramed.com) accompanies a subsequent version.

Acknowledgments

We thank the producers and maintainers of the public resources used in this work: GEO (GSE40279, GSE210255), DisGeNET, MSigDB (GO Biological Process gene sets), the Human Protein Atlas, and the STITCH chemical-protein interaction database.

References

1. ^{a, b, c, d, e}Gross B, Ehlert J, Gladyshev VN, Loscalzo J, Barabasi AL (2026). "Network-Driven Discovery of Repurposable Drugs Targeting Hallmarks of Aging." *Nat Aging*. 6(7):1516–1531. doi:[10.1038/s43587-026-01161-8](https://doi.org/10.1038/s43587-026-01161-8).

2. ^{a, b}Lopez-Otin C, Blasco MA, Partridge L, Serrano M, Kroemer G (2023). "Hallmarks of Aging: An Expanding Universe." *Cell*. **186**(2):243–278. doi:[10.1016/j.cell.2022.11.001](https://doi.org/10.1016/j.cell.2022.11.001).
3. ^{a, b}Lopez-Otin C, Blasco MA, Partridge L, Serrano M, Kroemer G (2013). "The Hallmarks of Aging." *Cell*. **153**(6):1194–1217. doi:[10.1016/j.cell.2013.05.039](https://doi.org/10.1016/j.cell.2013.05.039).
4. ^{a, b, c}Wang Z, Zhang O, Wang J, Jiang Y, Ma ZC, Guo M, Al Dajani S, Eames A, Moldakozhayev A, Poganik JR, Moqri M, Gobel D, Zhai R, Isakova A, Wu Y, Yin Z, Cong L, Ruohola-Baker H, Zou J, Snyder MP, Gladyshev VN, Wyss-Coray T, Ying K (2026). "A World Model Simulates the Latent Dynamics of Human Health." *medRxiv*. doi:[10.64898/2026.09.19.26363460](https://doi.org/10.64898/2026.09.19.26363460).
5. ^{a, b}Xiong J (2026). "World Models for Biomedicine: A Steerability Framework." *Preprints*. doi:[10.20944/preprints202605.0366.v1](https://doi.org/10.20944/preprints202605.0366.v1).
6. ^ΔXiong J (2026). "SteeraMed: A Biomedical World Model for N-of-1 Intervention Reasoning Across Chronic Diseases and Aging." *Preprints*. doi:[10.20944/preprints202605.1578.v1](https://doi.org/10.20944/preprints202605.1578.v1).
7. ^ΔXiong J, Xia Q (2026). "Toward a Self-Learning AI Agent for Drug Repurposing: Building Human-Scale Representations for Virtual Patients." *Preprints*. doi:[10.20944/preprints202608.0998.v1](https://doi.org/10.20944/preprints202608.0998.v1).
8. ^ΔLiang F, Quan Y, Wu A, Chen Y, Xu R, Zhu Y, Xiong J (2021). "Insulin-Resistance and Depression Cohort Data Mining to Identify Nutraceutical Related DNA Methylation Biomarker for Type 2 Diabetes." *Genes & Diseases*. **8**(5):669–676. doi:[10.1016/j.gendis.2020.01.013](https://doi.org/10.1016/j.gendis.2020.01.013).
9. ^ΔHannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sada S, Klotzle B, Bibikova M, Fan JB, Gao Y, Deconde R, Chen M, Rajapakse I, Friend S, Ideker T, Zhang K (2013). "Genome-Wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates." *Mol Cell*. **49**(2):359–367. doi:[10.1016/j.molcel.2012.10.016](https://doi.org/10.1016/j.molcel.2012.10.016).
10. ^ΔShang L, Zhao W, Wang YZ, Li Z, Choi JJ, Kho M, Mosley TH, Kardina SLR, Smith JA, Zhou X (2023). "meQTL Mapping in the GENOA Study Reveals Genetic Determinants of DNA Methylation in African Americans." *Nat Commun*. **14**(1). doi:[10.1038/s41467-023-37961-4](https://doi.org/10.1038/s41467-023-37961-4).
11. ^ΔBenjamini Y, Hochberg Y (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *J R Stat Soc Series B*. **57**(1):289–300. doi:[10.1111/j.2517-6161.1995.tb02031.x](https://doi.org/10.1111/j.2517-6161.1995.tb02031.x).
12. ^ΔPinero J, Ramirez-Anguila JM, Sauch-Pitarch J, Ronzano F, Centeno E, Sanz F, Furlong LI (2020). "The DisGeNET Knowledge Platform for Disease Genomics: 2019 Update." *Nucleic Acids Res*. **48**. doi:[10.1093/nar/gkz1021](https://doi.org/10.1093/nar/gkz1021).
13. ^ΔLiberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P (2015). "The Molecular Signatures Database (MSigDB) Hallmark Gene Set Collection." *Cell Syst*. **1**(6):417–425. doi:[10.1016/j.cels.2015.12.004](https://doi.org/10.1016/j.cels.2015.12.004).

14. [△]Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szgyarto CA, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist PH, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten Marica, von Feilitzen K, Forsberg M, Persson L, Johansson F, Zwahlen M, von Heijne G, Nielsen J, Pontén F (2015). "Tissue-Based Map of the Human Proteome." *Science*. **347**(6220):1260419. doi:[10.1126/science.1260419](https://doi.org/10.1126/science.1260419).
15. [△]Kuhn M, Szklarczyk D, Pletscher-Frankild S, Blicher TH, von Mering C, Jensen LJ, Bork P (2014). "STITCH 4: Integration of Protein-Chemical Interactions with User Data." *Nucl Acids Res*. **42**(D1):D401–D407. doi:[10.1093/nar/gkt1207](https://doi.org/10.1093/nar/gkt1207).
16. [△]Brin S, Page L (1998). "The Anatomy of a Large-Scale Hypertextual Web Search Engine." *Comput Networks ISDN Syst*. **30**:107–117. doi:[10.1016/S0169-7552\(98\)00110-X](https://doi.org/10.1016/S0169-7552(98)00110-X).
17. [△]Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008). "Fast Unfolding of Communities in Large Networks." *J Stat Mech*. **2008**(10):P10008. doi:[10.1088/1742-5468/2008/10/p10008](https://doi.org/10.1088/1742-5468/2008/10/p10008).
18. [△]World Health Organization (2020). "Decade of Healthy Ageing 2021–2030: Plan of Action." World Health Organization.
19. [△]Wang Y, Xiao S, Liu B, Jiang R, Liu Y, Hang Y, Chen L, Chen R, Vitiello MV, Bennett D, Wang B, Lv J, Yu C, Haslam DE, Zheng Q, Gerszten RE, Bao Y, Shi J, Xie J, Lu L, Li L, van Duijn CM, Wang DD, Chen Z, Chan AT (2025). "Organ-Specific Proteomic Aging Clocks Predict Disease and Longevity Across Diverse Populations." *Nat Aging*. **6**(1):162–180. doi:[10.1038/s43587-025-01016-8](https://doi.org/10.1038/s43587-025-01016-8).
20. [△]Tong H, Guo X, Teschendorff AE (2024). "Cell-Type Specific Epigenetic Clocks to Quantify Biological Age at Cell-Type Resolution." *Aging*. **16**(22):13452–13504. doi:[10.18632/aging.206184](https://doi.org/10.18632/aging.206184).
21. [△]Hao DC, Xiao PG (2014). "Network Pharmacology: A Rosetta Stone for Traditional Chinese Medicine." *Drug Dev Res*. **75**(5):299–312. doi:[10.1002/ddr.21214](https://doi.org/10.1002/ddr.21214).
22. [△]Xiong J (2026). "World Models for Biomedicine: Prediction Is the Means, Selection Is the End." *Preprints*. doi:[10.20944/P9-DOI-PENDING](https://doi.org/10.20944/P9-DOI-PENDING).
23. [△]Xiong J, the GHP Initiative Consortium (2026). "When One Person Is the Trial: Toward Good Healthspan Practice (GHP), a Draft Good-Practice Guide for N-of-1 Evidence in Longevity Medicine." *Preprints*. doi:[10.20944/preprints202609.1682.v1](https://doi.org/10.20944/preprints202609.1682.v1).
24. [△]Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, Jassal B, Jupe S, Matthews L, May B, Palatnik S, Rothfels K, Shamovsky V, Song H, Williams M, Birney E, Hermjakob H, Stein L, D'Eustachio P (2014). "The Reactome Pathway Knowledgebase." *Nucleic Acids Res*. **42**(D1):D472–D477. doi:[10.1093/nar/gkt1102](https://doi.org/10.1093/nar/gkt1102).

25. [△]Kanehisa M, Goto S (2000). "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Res.* **28**(1):27–30. doi:[10.1093/nar/28.1.27](https://doi.org/10.1093/nar/28.1.27).
26. [△]Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, von Mering C (2019). "STRING v11: Protein–Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets." *Nucleic Acids Res.* **47**(D1):D607–D613. doi:[10.1093/nar/gky1131](https://doi.org/10.1093/nar/gky1131).
27. [△]Wilcoxon F (1945). "Individual Comparisons by Ranking Methods." *Biometrics Bull.* **1**(6):80–83. doi:[10.2307/3001968](https://doi.org/10.2307/3001968).
28. [△]Barardo D, Thornton D, Thoppil H, Walsh M, Sharifi S, Ferreira S, Anžič A, Fernandes M, Monteiro P, Grum T, Cordeiro R, De-Souza EA, Budovsky A, Araujo N, Gruber J, Petrascheck M, Fraifeld VE, Zhavoronkov A, Moskalev A, de Magalhães JP (2017). "The DrugAge Database of Aging-Related Drugs." *Aging Cell.* **16**(3):594–597. doi:[10.1111/acer.12585](https://doi.org/10.1111/acer.12585).
29. [△]Rafikova E, Nemirovich-Danchenko N, Ogmen A, Parfenenkova A, Velikanova A, Tikhonov S, Peshkin L, Rafikov K, Spiridonova O, Belova Y, Glinin T, Egorova A, Batin M (2024). "Open Genes—a New Comprehensive Database of Human Genes Associated with Aging and Longevity." *Nucleic Acids Res.* **52**(D1):D950–D962. doi:[10.1093/nar/gkad712](https://doi.org/10.1093/nar/gkad712).

Supplementary data: available at <https://doi.org/10.32388/A1NTLL>

Declarations

Funding: This research was conducted and funded by DeepoMe Inc. as in-house scientific research.

Potential competing interests: The author J.X. is the founder of DeepoMe Inc. (Beijing, China), which develops DNA methylation-based health assessments and related products and services. This research was conducted and funded by DeepoMe Inc. as in-house scientific research. No other financial or non-financial competing interests are declared.