ARTICLE OPEN

iS2C2: a cointelligent platform for mechanistic discovery of disease cellular crosstalk

Jianting Sheng[1,2]✉, Ju Young Ahn[1,3,4], Li Yang[1,2]✉, Zhihao Wan[1,2], Shaohua Qi[1,2], Xiaohui Yu[1,2], Zhan Xu[5,6,7], Yuliang Cao[1], Matthew Vasquez[1], Amna Irfan[1], Yuanyuan Zhu[1,8], Hong Zhao[1,2], Zheng Yin[1,2], Ying Zhu[9], Yunfeng Ding[5,6,7], Alireza Faridar[10], Lin Wang[1,2], Fengshuo Liu[5,6,7], Hongxia Wang[11], Zhigang Ji[12], Dongxue Mao[7], Michael Chan[1,2], Daniel Kermany[1,3,4], Wenjuan Dong[1,2], Doo Yeon Kim[13,14], Xiang H.-F. Zhang[5,6,7] and Stephen T. C. Wong[1,2,3,4,5,6,7,8,10,15]✉

Large language models (LLMs) have demonstrated impressive capabilities in summarization, reasoning, and content generation, yet their inability to directly interpret large-scale omics data has limited their utility in data-driven hypothesis generation—particularly in mechanism discovery that demands the integration and interpretation of multimodal datasets, heterogeneous models, and deep domain expertise. Conversely, traditional computational algorithms excel at quantitative analysis of omics data but often rely heavily on labor-intensive, expert-driven interpretation to extract biologically meaningful insights. Here, we introduce (*cointelligent single-cell spatial cell–cell communication*: iS2C2), a novel cointelligent platform that synergizes mathematically rigorous computational algorithms with the contextual reasoning capabilities of LLMs to automatically generate biologically interpretable hypotheses from single-cell RNA-seq and spatial transcriptomics data. The iS2C2 platform incorporates a transparent and reproducible cell–cell communication analysis pipeline built upon mathematically rigorous algorithms designed to enhance interpretability for integration with LLMs that contextualize algorithmic outputs or predictions using domain-specific knowledge and literature-derived evidence. When applied to Alzheimer's disease and cancer datasets, iS2C2 generated accurate, reproducible, and expert-validated hypotheses, unveiling previously unrecognized signaling pathways and mechanistic insights in disease microenvironments. This cointelligent approach bridges the gap between structured computational analysis and generative reasoning, heralding a paradigm shift toward fully automated, interpretable biological discovery and advancing the frontiers of next-generation precision medicine and systems biology.

Signal Transduction and Targeted Therapy (2026)11:172 ; https://doi.org/10.1038/s41392-026-02691-8

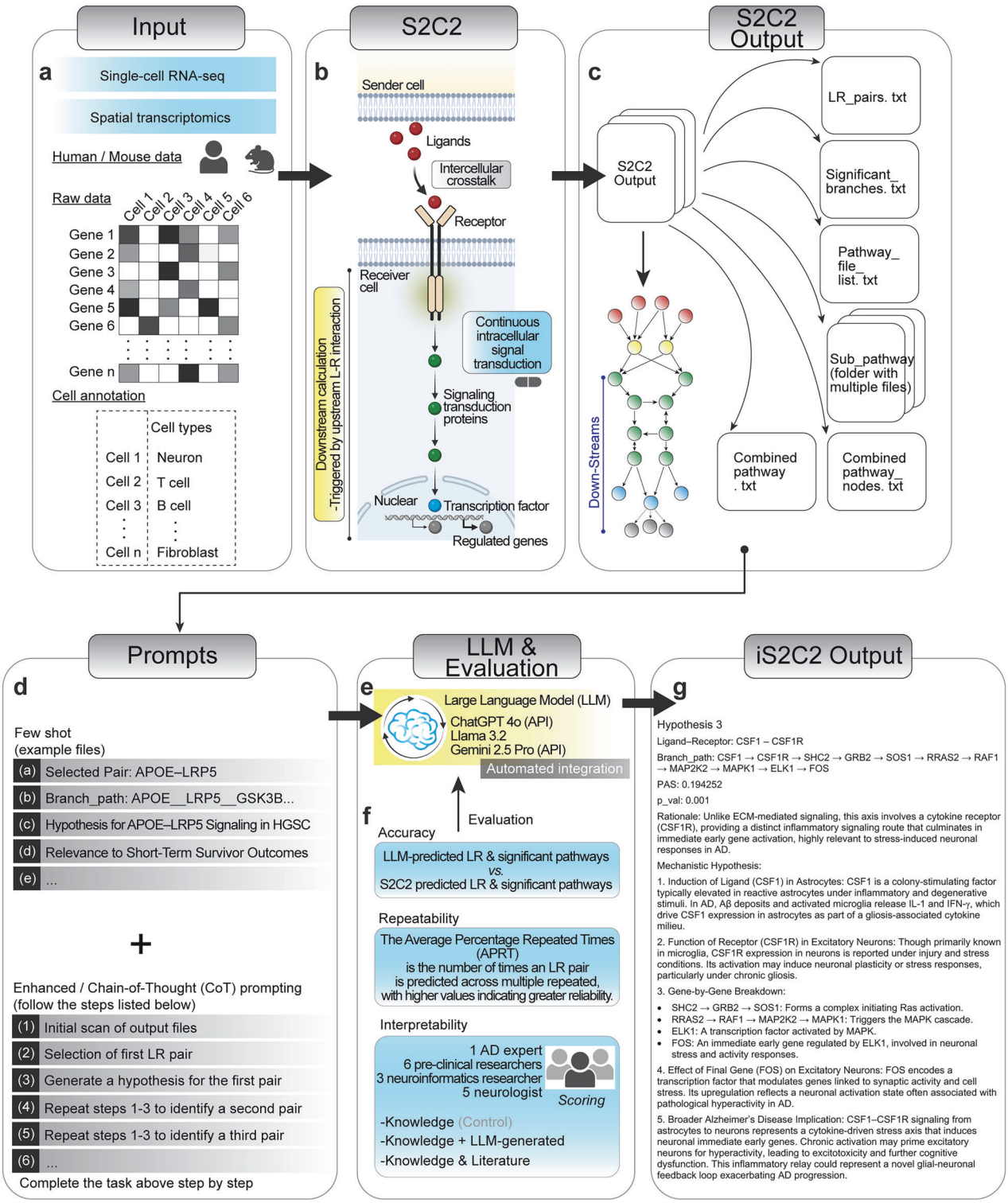
INTRODUCTION

Generating robust and experimentally testable hypotheses in biological research requires the integration of high-dimensional omics data, sophisticated computational modeling, and deep domain expertise. However, this integration remains challenging for bioinformaticians, biologists, clinicians, and data scientists due to the complexity of biological data and the communication barriers that often exist between computational and experimental disciplines. Recent advances in large language models (LLMs) have significantly transformed knowledge discovery and dissemination in biology and medicine, enabling powerful applications in knowledge sharing, content summarization, code generation, translation, and education.[1–4] Despite these capabilities, LLMs lack the intrinsic ability to directly interpret large-scale omics data or derive mechanistic insights from high-dimensional biological measurements.[5,6] In parallel, traditional computational methods have made substantial progress in analyzing omics datasets, including identifying biomarkers,[7,8] reconstructing cell trajectories,[9–11] and modeling cell–cell communication networks.[12–15] These methods can effectively extract patterns from complex datasets, yet their outputs are often mathematically intricate and difficult for researchers to interpret in a mechanistic context. As a result, translating computational findings into actionable biological hypotheses remains a major challenge.

One particularly important area where this challenge is evident is the study of cell–cell communication, which plays a fundamental role in regulating tissue homeostasis and disease progression. Increasing evidence indicates that dysregulated intercellular signaling drives pathological processes across diverse

[1]Systems Medicine and Bioengineering Department, Houston Methodist Neal Cancer Center, Houston Methodist Hospital, Houston, TX, USA; [2]Ting Tsung & Wei Fong Chao Center for BRAIN, Houston Methodist Hospital, Houston, TX, USA; [3]Texas A&M University College of Medicine, Bryan, TX, USA; [4]Department of Biomedical Engineering, Texas A&M University, College Station, TX, USA; [5]Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, TX, USA; [6]Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, TX, USA; [7]Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX, USA; [8]Department of Bioengineering, Rice University, Houston, TX, USA; [9]Department of Genomic Medicine, University of Texas M.D. Anderson Cancer Center, Austin, TX, USA; [10]Departments of Radiology, Medicine, and Urology, Houston Methodist Hospital, Houston, TX, USA; [11]Department of Laboratory Medicine, University of Texas M.D. Anderson Cancer Center, Austin, TX, USA; [12]Med-Surg units, Ben Taub Hospital, Austin, TX, USA; [13]Genetics and Aging Research Unit, Department of Neurology, Mass General Institute for Neurodegenerative Disease, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA; [14]McCance Center for Brain Health, Massachusetts General Hospital, Boston, MA, USA and [15]Departments of Radiology, Neuroscience, Pathology and Laboratory Medicine, Weill Cornell Medicine, Houston, TX, USA
Correspondence: Jianting Sheng (jsheng@houstonmethodist.org) or Li Yang (lyang@houstonmethodist.org) or Stephen T. C. Wong (stwong@houstonmethodist.org)
These authors contributed equally: Jianting Sheng, Ju Young Ahn, Li Yang, Zhihao Wan

Received: 6 May 2025 Revised: 6 February 2026 Accepted: 15 March 2026
Published online: 11 May 2026



biological contexts, including neurodegenerative disorders, cancer progression, and metastatic dissemination.[16–18] In these systems, ligand–receptor (*L–R*) interactions trigger downstream signaling cascades that reshape cellular states, remodel tissue microenvironments, and propagate pathological signals across interacting cell types. Advances in single-cell and spatial transcriptomics have provided unprecedented opportunities to study these communication networks at cellular resolution.[12,13] Consequently, several computational approaches have been developed to infer cell–cell communication from transcriptomic data.[12–19] Some existing methods focus primarily on identifying *L–R* interactions,[14,17] while others extend the analysis by incorporating downstream targets[15] or applying matrix decomposition approaches to approximate signaling dynamics.[19] Although these strategies have improved our ability to characterize intercellular signaling landscapes, they often provide limited mechanistic resolution. In particular, many approaches lack explicit modeling of downstream regulatory pathways or fail to distinguish converging signaling routes,

Fig. 1 Workflow of iS2C2: a cointelligent platform for cell–cell communication and hypothesis prediction. **a** The platform begins with input from single-cell or spatial transcriptomic datasets, applicable to both human and mouse samples. Raw gene expression matrices and cell type annotations were used as input for downstream analysis. **b** The S2C2 algorithm identifies intercellular crosstalk by computing ligand–receptor (*L–R*) interactions and downstream signal propagation in a continuous manner. **c** S2C2 generates structured output files summarizing predicted ligand–receptor interactions, downstream signaling branches, and pathway context. • LR_pairs.txt—contains differential expression, specificity scores, and enrichment levels for each ligand–receptor pair. • significant_branches.txt—lists statistically significant downstream signaling branches with associated weights and *p* values. • pathway_file_list.txt—provides pathway names, significance *p* values, and links to corresponding node/edge files. • sub_pathway/*(folder)*—Includes gene-level node files for each significant pathway. • combined_pathway_-nodes.txt—merges all pathway node files, annotating each gene's role (e.g., ligand, receptor, transcription factor) and associated pathways. • combined_pathway.txt—integrates interaction relationships across all pathways, capturing directionality and interaction types between gene nodes. **d** These outputs are converted into prompts: few-shot analogical prompting (top), which provides example files such as ligand–receptor pairs, signaling branches, hypotheses, and clinical relevance; and enhanced chain-of-thought (CoT) prompting (bottom), which guides the large language model (LLM) step-by-step—from scanning output files to selecting key *L–R* pairs and generating hypotheses. **e** The S2C2 framework enables automated integration with multiple large language models (LLMs), including Chatbox (ChatGPT via API), Llama 3.2, and Gemini 2.5 Pro (see Supplementary Fig. 16a for details), allowing seamless downstream hypothesis generation without manual intervention. **f** The LLM outputs are evaluated based on three criteria: accuracy (agreement between LLM and S2C2 results), repeatability (average percentage repeated times, APRT), and interpretability, assessed by a panel of domain experts, including AD specialists, neuroinformatics researchers, and neurologists. **g** The final output of the integrated S2C2–LLM framework (iS2C2) is a ranked set of ligand–receptor pairs with accompanying mechanistic hypotheses, providing candidate signaling axes and testable hypotheses in disease contexts such as Alzheimer's disease (AD)

making it difficult to connect specific *L–R* interactions to concrete functional consequences.

These limitations present a major obstacle for integrating computational models with LLM-based reasoning systems. For LLMs to generate reliable and biologically meaningful hypotheses, the input information must be structured, interpretable, and mechanistically grounded. However, when computational outputs lack explicit mechanistic pathways or clear quantitative representations of signaling dynamics, the resulting information becomes underconstrained for LLM interpretation. This can increase the likelihood of hallucination, inconsistent reasoning, and reduced reproducibility in hypothesis generation.[20–23] Therefore, improving the interpretability and mechanistic transparency of computational models is a critical prerequisite for enabling effective collaboration between algorithmic analysis and LLM-driven reasoning. To address this challenge, we sought to develop computational frameworks that are inherently LLM-adapted, meaning that their outputs are structured in a way that can be readily interpreted, summarized, and reasoned upon by both researchers and AI systems. Specifically, we aimed to enhance existing cell–cell communication modeling strategies by incorporating multilayer signaling pathways and quantitative pathway activity scoring, thereby providing a more detailed representation of how *L–R* interactions propagate through intracellular regulatory networks.

Here we present iS2C2 (cointelligent single-cell spatial cell–cell communication), a cointelligent computational platform that integrates interpretable signaling algorithms with LLM-assisted reasoning to facilitate hypothesis generation in complex biological systems. iS2C2 expands traditional cell–cell communication models by explicitly mapping multilayer downstream signaling pathways and assigning quantitative Pathway Activity Scores (PASs) to prioritize both intercellular and intracellular signaling events. These structured outputs are then integrated with LLM-enhanced interpretative capabilities through tailored prompt engineering, enabling synergistic interactions between quantitative algorithms and AI-driven reasoning. To evaluate the robustness of iS2C2, we compared it with baseline frameworks constructed by pairing existing crosstalk inference methods with LLMs across four evaluation dimensions: sensitivity to perturbation, accuracy and faithfulness, repeatability, and interpretability. We further demonstrate the utility of iS2C2 by applying it to uncover previously unrecognized cell–cell communication pathways in Alzheimer's disease (AD) and cancer. Experimental validation supports the biological relevance of the identified mechanisms, illustrating how the platform can enable transparent, interpretable, and automated hypothesis generation. Together, these results highlight the potential of cointelligent AI systems to transform biological discovery by bridging computational modeling with human and machine-assisted reasoning.

RESULTS

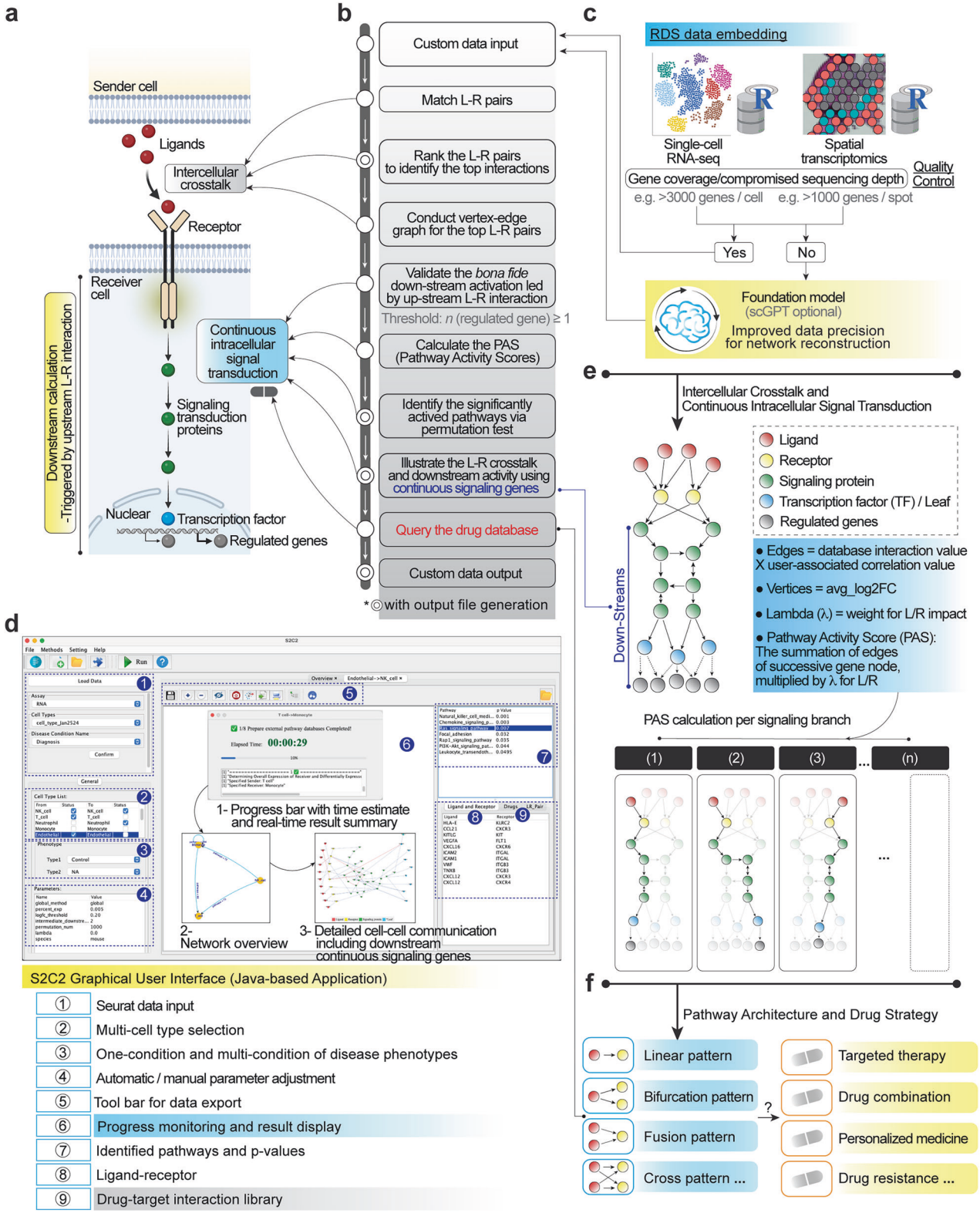
Design of the iS2C2 cointelligent platform

iS2C2 (https://github.com/methodistsmab/iS2C2) is engineered to integrate a cellular crosstalk model called single-cell spatial cell–cell communication (S2C2), with an LLM for seamless, robust hypothesis generation (Fig. 1). To achieve this, we implemented targeted design improvements in both components, ensuring seamless communication and efficient processing.

Cellular crosstalk model (S2C2) design. The primary goal of the cellular crosstalk model is to generate detailed predictions of *L–R* interactions and their associated downstream, knowledge-based pathways by integrating single-cell RNA-seq and spatial transcriptomic data from human and mouse tissues (Fig. 1a, b). These predictions are structured for seamless interpretation by the LLM (Fig. 1c). To maximize their translational value, we introduce scoring algorithms to quantify the significance of individual *L–R* pairs and pathways, enabling the LLM to prioritize high-impact targets for robust hypothesis generation. A foundation model-based gene expansion strategy overcomes low-coverage sequencing limits, substantially improving prediction capacity.

LLM design. To maximize the effectiveness of the LLM, we guided it with system prompts that clearly define its task. We provided few-shot learning or analogical reasoning to illustrate the desired hypothesis format (Fig. 1d). To further improve robustness evaluated by accuracy and reproducibility, we incorporate chain of thought (CoT) and few-shot approaches, which encourage a step-by-step reasoning process that aligns with human logical inference during hypothesis formulation (Fig. 1e). The quality of the generated hypotheses is rigorously assessed by fifteen domain experts through a structured questionnaire that evaluates interpretability in addition to accuracy and reproducibility (Fig. 1f, g), a strategy that has been successfully applied in similar contexts for drug repurposing.[24]

To systematically evaluate iS2C2 performance, we defined robustness along four complementary axes: sensitivity to perturbation, accuracy/faithfulness, repeatability, and interpretability. We compared each axis against baseline crosstalk algorithms (LIANA+, NicheNet, etc.) or baseline LLM frameworks (crosstalk algorithm + LLM).



1. Sensitivity to perturbation of the crosstalk model: this axis evaluates how strongly pathway-level *L–R* predictions respond to biologically meaningful changes in gene expression. Using in silico perturbation experiments, we manually up- or downregulated specific *L–R* pairs and quantified how S2C2, LIANA+, and NicheNet reflect these changes in both *L–R* predictions and downstream ligand–target outputs. Greater sensitivity indicates better alignment with biologically meaningful *L–R* perturbations and stronger capability to capture changes in cell–cell communication.
2. Accuracy/faithfulness of the LLM: accuracy measures LLM faithfulness to its own crosstalk backbone, not biological truth. For each sender–receiver pair, we determined the

Fig. 2 Workflow, user interface, and advancements of S2C2. **a** Schematic diagram of S2C2 highlighting its functionality to depict intracellular continuous signal transduction (pathway branches) initiated by upstream ligand–receptor interactions. **b** The general workflow of S2C2 includes intracellular continuous signaling pathways and intercellular crosstalk between ligands and receptors. **c** S2C2 enables direct embedding of Seurat data from scRNA-seq (over 3000 genes per cell) and spatial transcriptomics (over 1000 genes per spot) when gene coverage meets quality control standards. Otherwise, a foundation model such as scGPT can be applied to expand gene coverage. One of the main challenges in modeling cellular crosstalk using scRNA-seq and ST data lies in the limited gene coverage and low gene expression detection within each cell and ST spot, which restricts the number of detectable genes and hampers the ability to discern significant gene fold changes (FCs). This limitation makes traditional statistical differential analysis, which relies on individual gene FC to assess pathway FC, particularly challenging. Additionally, studies often need to balance sequencing depth across all cell types, including immune or rare cell types where specific genes are expressed at very low levels, further complicating deep sequencing feasibility. **d** The S2C2 user interface (UI) developed in Java offers a range of advanced features to facilitate comprehensive data analysis and visualization: (1) Direct input of Seurat data, including scRNA-seq and spatial transcriptomics data. (2) Multicell type selection. (3) Support for both single-condition and multicondition selection for disease phenotypes. (4) Options for automatic and manual parameter adjustment to fine-tune analyses. (5) Direct data output. (6) Estimation of computational runtime and visualization of outcomes in real time. (7) Visualization of detected pathways and their associated *P* values for easy interpretation. (8) Display of predicted ligand–receptor pairs. (9) Integration of a drug library to identify and target pathways or genes. **e** S2C2 enables the illustration of intact pathways, including ligands, receptors, signaling proteins, transcription factors, and regulated genes, based on the calculation of the Pathway Activity Score (PAS) for each branch. **f** The integration of pathway patterns and drug libraries in S2C2 enhances its capacity to perform multiple functions, such as guiding drug administration and dissecting molecular mechanisms

proportion of LLM-generated top-ranked *L–R* or pathway predictions that matched the backbone model's outputs (iS2C2 vs. S2C2; LIANA+LLM vs. LIANA; NicheNet+LLM vs. NicheNet). Outputs proposed by the LLM but absent from the backbone constitute hallucinations, and their frequency (1—accuracy) defines the hallucination rate. Higher faithfulness reflects better preservation of backbone predictions and fewer LLM-introduced distortions.

3. Repeatability: repeatability measures the stability of LLM outputs across independent runs under identical settings. We quantified this using the average percentage repeated times (APRT), where higher values indicate lower stochastic variability. Greater repeatability reflects more stable output patterns and reduces stochastic variation across runs.
4. Interpretability: interpretability is assessed through expert review. Fifteen domain experts evaluated LLM-generated hypotheses across multiple task designs, scoring each method on confidence, ranking, and disease relevance. These metrics reflect how understandable, plausible, and clinically meaningful the generated hypotheses are to human experts. Higher interpretability indicates better alignment with expert reasoning and greater practical utility.

Existing methods incorporate design choices that may influence specific aspects of robustness. Specifically, LIANA+ employs ensemble ranking, which may improve repeatability; S2C2 and NicheNet include downstream signaling modeling, which may enhance interpretability; and iS2C2(S2C2+LLM) uses prompt engineering and structured reasoning (CoT, few-shot examples) to improve LLM faithfulness, repeatability, and interpretability. We assessed robustness empirically across the four axes to provide a transparent and method-agnostic comparison.

Development of an LLM-interpretable algorithm (S2C2) for single-cell and spatial crosstalk prediction

Directly using the outputs of current crosstalk algorithms with LLMs for hypothesis generation presents two major challenges: (1) predictions limited solely to *L–R* interactions may ignore the downstream pathway activity, and (2) direct integration with LLMs may suffer from hallucination and reproducibility issues.[25,26]

To address these limitations, we developed S2C2, a novel algorithm for precise mapping of intercellular and intracellular signal transduction across datasets with diverse gene coverage. Its user-friendly graphical interface (https://github.com/methodistsmab/S2C2/) supports parallel, multicellular crosstalk analysis. Specifically, S2C2:

1. introduce pathway branch mapping to perform knowledge-driven, downstream mapping of continuous intracellular signal transduction, identifying convergent pathway patterns or disease mechanisms originating from diverse cell types and *L–R* interactions, and vice versa (Fig. 2a, b),
2. incorporates a deep learning-based foundational model that enhances the performance of intercellular crosstalk predictions using datasets from restricted sequencing depth and gene coverage, common in single-cell and ST sequencing (Fig. 2c),
3. enables parallel multicellular crosstalk analysis across multiple CPU cores, allowing simultaneous investigation of feedback loops in cell–cell communication, and
4. features a graphical user interface (GUI) (Fig. 2d) for interactive analysis and visualization.

Furthermore, S2C2 seamlessly integrates with the standard R Seurat workflow[27] and the Linux command-line interface (CLI) (Supplementary Fig. 1). These advancements could enhance data interpretability, prioritization of predictions of *L–R* pairs, identification of downstream intracellular pathways and targets, and user experience.

S2C2 curates and integrates multiple databases, including an *L–R* interaction database,[13] several signaling pathway databases, including Kyoto Encyclopedia of Genes and Genomes (KEGG, updated Jan 2024) and Ingenuity Pathway Analysis (IPA, version 22.0.2), and a transcription factor (TF) database,[13] to comprehensively predict intercellular crosstalk of *L–R* interactions and their stratified downstream biological pathway branches (Supplementary Fig. 2). Specifically, the CCCExplorer *L–R* database provides curated *L–R* interactions that serve as the starting edges for S2C2, whereas the KEGG and IPA supply the downstream intracellular signaling pathways that connect receptors to signaling intermediates and transcriptional regulators cataloged in the TF–target database. The TF–target database is used both to annotate transcription factors and to support cascading communication prediction across more than two cell types. Together, these resources allow S2C2 to model both *L–R* interactions and intracellular signaling propagation across diverse cell types.

Additionally, S2C2 orchestrates a drug discovery function into the GUI, leveraging the DrugBank database[28] and extendable computational approaches such as the Connectivity Map[29] (Supplementary Fig. 3). This innovation can enhance therapeutic target identification, improve functional interpretation of key genes for personalized therapeutic strategies, and accelerate the translation of findings into clinical applications.

S2C2 prioritizes *L–R* pairs based on metrics such as average fold change, percentage of expressed genes, and *P* value generated from permutation tests (Supplementary Fig. 4 and Supplementary Note 1A). Pairwise analysis was conducted to rank *L–R* pairs based

on their enrichment scores (Supplementary Fig. 4). Unlike existing crosstalk algorithms that rely solely on absolute gene expression for *L–R* prediction, S2C2 enables cross-analysis across various cell types and disease states, such as comparing multiple cell types within a healthy condition, analyzing the same cell type across different disease states, and examining different cell types across various disease conditions (Fig. 2d).

To further enhance interpretability for LLM, S2C2 defines pathway branches (Supplementary Fig. 5) and calculates a PAS that incorporates the weighted contributions of both *L–R* interaction pairs and the strength of downstream pathways (Supplementary Fig. 6a, b and Supplementary Note 1B). A pathway branch is defined as a subgraph within a validated biological pathway network linking *L–R* interactions to terminal nodes, i.e., genes with no outgoing edges under the directed graph generated from KEGG/IPA pathway topology. In graph-theoretical terms, each gene within a pathway branch has one incoming and one outgoing edge, except for the start node (e.g., the ligand) and the end node (e.g., the transcription factor or their targets) (Fig. 2e). A biological pathway can contain multiple pathway branches, each exhibiting distinct activation patterns that differentiate cell types under varying conditions (Fig. 2e). S2C2 determines pathway significance using permutation tests for the outcome display (Supplementary Fig. 6c, d). These diverse pathway patterns can offer novel insights and therapeutic intervention strategies (Fig. 2f).

The stepwise construction of pathway branches and the subsequent branch-based modeling workflow are outlined below:

1. Build an integrated directed graph using *L–R* interactions from CCCExplorer and downstream signaling edges from KEGG and IPA.
2. For each *L–R* pair, identify all shortest directed paths from the receptor node to terminal nodes, i.e., genes with no outgoing edges in the graph, annotated as either transcription factors or leaf nodes (e.g., TF target genes).
3. Define each such directed path as a pathway branch.
4. Retain branches that contain at least two intermediate genes between the receptor and the terminal node.
5. Compute a PAS by aggregating weighted log2-fold changes for all genes along the branch, accounting for interaction direction and type.
6. Perform permutation tests to identify branches whose PAS values are significantly higher (or lower) than expected under random gene label permutations.

In addition, the hyperparameter λ serves as a weighting factor to balance the contributions of *L–R* interactions and downstream branch predictions when calculating the PAS score, optimizing the trade-off between false-positive and false-negative rates (Supplementary Fig. 7). We also explored the impact of varying hyperparameters, including fold-change thresholds and percentage of expressed cells, on model predictions (Supplementary Figs. 8 and 9), revealing consistency with biological expectations.

Comparative analysis and perturbation sensitivity of S2C2

To assess concordance among different crosstalk inference algorithms, we applied S2C2, NicheNet,[15] and various LIANA +-based[17,30–34] methods to Tsai's AD dataset.[35] The number of predicted *L–R* pairs between major brain cell types varied substantially across algorithms (Supplementary Fig. 10a). S2C2 generally produced a moderate number of predicted interactions for most cell-type pairs (Supplementary Fig. 10a). We then constructed Venn diagrams for each cell-type pair, comparing predicted *L–R* pairs identified by S2C2, LIANA_CellChat, LIANA_rank_aggregate, and NicheNet (Supplementary Fig. 10b). The results reveal substantial variability in the degree of overlap among algorithms. Across most combinations, only a small fraction of *L–R* pairs were shared by all four methods, with each algorithm also identifying a considerable number of unique interactions. Notably, S2C2 and NicheNet exhibited lower overlap with LIANA+-based methods, likely reflecting differences in whether downstream pathway information was incorporated into the prediction models (Supplementary Fig. 10b). These findings demonstrate that *L–R* inference is highly method-dependent.

We next evaluated the sensitivity to the perturbation axis of robustness by manually altering the expression of selected *L–R* pairs. In silico validation demonstrated that S2C2 reliably identified pathways when ligand and pathway genes were manually amplified and failed to detect specific pathways when their *L–R* interactions were artificially reduced (Supplementary Fig. 11a–c). These bidirectional perturbations confirm that the sensitivity of S2C2 in pathway prediction is tightly associated with the expression levels of *L–R* pairs and their downstream genes. When MAPK *L–R* pairs were upregulated, nearly all algorithms showed increased detection of MAPK *L–R* interactions, except for NicheNet (Supplementary Fig. 12a). NicheNet also showed sensitivity to changes at the downstream ligand–target level, as S2C2 does, although they utilize different databases and algorithms, resulting in distinct downstream predictions (Supplementary Fig. 12b and Supplementary Note 1C). Upon downregulation of RAS *L–R* pairs, S2C2 displayed a more pronounced sensitivity for both reduced *L–R* and downstream target predictions (Supplementary Fig. 12c, d).

In addition to Tsai's AD dataset, we further evaluated sensitivity to perturbation of S2C2 and other existing crosstalk algorithms using two non-brain immune datasets from the Tumor Immune Cell Atlas (TICAtlas),[36] focusing on M2 TAM → NK cell interactions in breast cancer (Supplementary Fig. 12e–h) and M2 TAM → cDC cell interactions in non-small cell lung cancer (NSCLC) (Supplementary Fig. 12i–l). Using the same perturbation strategy as the AD dataset, we upregulated NF-κB pathway ligands/receptors and downregulated RAP1 pathway components in breast cancer and upregulated MAPK ligands/receptors and downregulated focal adhesion components in NSCLC; we then compared the sensitivity of S2C2, nine LIANA+ methods, and NicheNet. All algorithms detected increased NF-κB signaling, and both S2C2 and NicheNet showed corresponding increases in downstream targets (Supplementary Fig. 12e, f). For RAP1 downregulation, S2C2 uniquely showed decreased *L–R* predictions, while downstream reductions were captured by both S2C2 and NicheNet (Supplementary Fig. 12g, h). Similar trends were identified for NSCLC comparisons (Supplementary Fig. S12i–l). These findings demonstrate that S2C2 responds consistently to perturbations across both neurodegeneration and cancer immune datasets, supporting its generalizability across diverse disease contexts and cell types.

Together, these findings demonstrate that S2C2 is both method-distinct and highly sensitive to biologically relevant perturbations in cell–cell communication networks, thus completing the sensitivity to perturbation axis within our robustness assessment based on different datasets.

S2C2 is compatible with low gene coverage sequencing data

Inference of cell–cell communication from single-cell and spatial transcriptomic data is constrained by limited gene coverage and sparse detection, which reduce the number of measurable transcripts and attenuate fold-change estimates (Fig. 2c). Such limitations are particularly problematic for low-abundance transcripts in certain immune and rare cell populations, where sequencing depth is often restricted by technical and cost considerations. Using an AD dataset,[35] we show that our algorithm performs robust predictions with reduced gene lists comprising 1000 genes that align with Nanostring's CosMX Human Universal Cell Characterization and Human 6K Discovery Panels compared with the original 18,000-gene coverage dataset (Supplementary Fig. 13a–d).

To further validate S2C2's sensitivity in capturing complete downstream branches under reduced coverage, we randomly selected 1000 genes from the AD scRNA-seq dataset[35] and used scGPT[37] to expand coverage back to 18,000 genes (Supplementary Tables 1 and 2, Supplementary Fig. 13e, f). By training a regression model on the scGPT embeddings, S2C2 recovered the majority of the top *L–R* pairs and a substantial number of downstream-associated genes and pathway branches, even under simulated gene dropout rates of 20%-80% (Supplementary Figs. 13e, f and 14). These findings indicate that S2C2 provides robust predictions, with sensitivity and specificity trade-offs that depend on the dropout rate configuration in the expansion model. The results showed expected patterns within realistic parameter spaces, underscoring the high interpretability and reliability of our model with respect to these hyperparameters.

Together, these results demonstrate that S2C2, particularly when integrated with gene coverage expansion via scGPT, enables imputation of missing gene expression values, recovery of otherwise undetected *L–R* pairs, and preservation of pathway-level predictions from incomplete transcriptomes. This approach maintains high prediction performance and biological interpretability even with incomplete transcriptomic information, thereby broadening its applicability to datasets generated with targeted panels or limited sequencing depth.

Design of an LLM for hypothesis generation from cellular crosstalk predictions

To translate predicted cell–cell interactions into testable biological hypotheses, we designed an LLM-driven workflow using GPT-4o (Fig. 3a), incorporating two tailored prompt-engineering strategies (Fig. 3b)[38,39]: (1) few-shot prompting and (2) chain-of-thought (CoT) prompting.

In the few-shot setup, the model is presented with representative, experimentally validated examples of cell–cell interactions curated from peer-reviewed literature[40–43] (Supplementary Note 2). These examples serve as contextual anchors to guide the model's reasoning within biologically realistic bounds. In parallel, CoT prompts stepwise reasoning by explicitly instructing GPT-4o to (1) identify relevant *L–R* pairs, (2) trace their downstream signaling cascades, and (3) assess their biological relevance in specific disease contexts.

Explicitly directing the model to "complete the task above step by step" promotes transparent intermediate reasoning and enhances both interpretability and predictive accuracy. All prompting strategies (Supplementary Note 2 and iS2C2 GitHub) are iteratively refined to improve clarity, task fidelity, and biological precision. This process involves systematic evaluation of the model's outputs: whenever ambiguity, factual inaccuracies, or incomplete biological logic are identified, the prompts are revised to include more granular step descriptions, clarified domain-specific terminology, and explicit biological constraints. This iterative optimization minimizes ambiguity, enhances interpretability, and links computational crosstalk predictions to testable disease-mechanism insights.

Robustness evaluation of the cointelligent platform iS2C2: accuracy and reproducibility axes

We next evaluated the accuracy/faithfulness axis of robustness for the cointelligent iS2C2 platform. We applied our cointelligent framework to eight sender–receiver cell type pairs, generated the top three *L–R* interactions, and repeated this procedure five times per pair. Here, accuracy does not measure biological correctness. Instead, it measures whether the LLM remains faithful to its own crosstalk model rather than hallucinating new interactions. For each sender–receiver pair, we compute the fraction of LLM-generated *L–R* pairs and pathways that also appear in that model's original predictions (iS2C2 vs. S2C2; LIANA+LLM vs. LIANA; NicheNet+LLM vs. NicheNet). One minus this value is the hallucination rate—the proportion of LLM-suggested interactions that do not exist in its own backbone model. The results showed that by employing enhanced prompting, the mean accuracy increased from 0.72 to 0.81 for zero-shot prompting ($p = 0.05$, one-tailed paired *t*-test) and from 0.68 to 0.78 for few-shot prompting ($p = 0.01$, one-tailed paired *t*-test) compared to simple prompting, which solely includes a description of the input format along with its intended purpose (Supplementary Fig. 15a). These findings indicate that utilizing multistep prompting can improve the accuracy of our hybrid platform, although few-shot prompting did not provide more accurate predictions compared to zero-shot prompting.

We then quantified the repeatability axis of robustness by measuring the consistency of *L–R* pair identification across repeated queries using identical prompts and input files. Specifically, we calculated the APRT as the proportion of *L–R* pairs appearing in at least 80% of repeated experiments. Enhanced prompting with a few-shot setup achieved the highest APRT (0.51), exceeding that of simple prompting with zero-shot (APRT = 0.46), simple prompting with few-shot (APRT = 0.28), and enhanced prompting with zero-shot (APRT = 0.41). Notably, enhanced prompting increased the APRT from 0.28 to 0.51 in the few-shot condition ($p = 0.02$, one-tailed paired *t*-test) (Supplementary Fig. 15b), although no other differences reached statistical significance—likely because of the small sample size and high variance among different cell types. For the top three hits, we further evaluated them using a larger sample size (10 repeats instead of 5 repeats for excitatory neuron to microglia and astrocyte interactions). The results showed that the top hits remained highly consistent across all repeats (10 of 10; see Supplementary Note 1D).

To systematically assess the technical performance and practical impact of LLM-based crosstalk prediction, we compared S2C2 with preexisting algorithms, including their LLM-augmented implementations (Supplementary Fig. 15c, d). Across multiple cell–cell interaction types, iS2C2 achieved consistently high accuracy and repeatability in predicting *L–R* pairs, performing comparably to other established methods in these experiments (Supplementary Fig. 15c, d and Supplementary Note 1E). Repeatability metrics further demonstrate the robustness and stability of S2C2 predictions across repeated runs. These findings demonstrate the technical feasibility of integrating LLMs into S2C2 for crosstalk inference.

To evaluate the performance of S2C2 when integrated with various LLMs, we systematically compared ChatGPT4.0, Llama3.2, Gemini 2.5 Pro, and KIMI K2 across multiple prompt strategies (simple zero-shot, simple few-shot, enhanced zero-shot, and enhanced few-shot). Llama3.2 served as an openly available local source, while Gemini and KIMI were selected for their accessibility via application programming interfaces (APIs) and Chatbox, which enabled either fully automated or partially manual integration with S2C2-driven *L–R* prediction (Supplementary Fig. 16a). The results demonstrate that S2C2 can be effectively combined with all tested LLMs, achieving consistently high accuracy and satisfactory repeatability in *L–R* pair prediction across diverse model sources and deployment modes (Supplementary Fig. 16b, c). These findings highlight the broad applicability and flexibility of the S2C2+LLM workflow, which achieves robust and reproducible performance in automated crosstalk inference and supports both locally deployed and API-based LLMs for seamless integration into hypothesis-driven analyses.

Robustness evaluation of the cointelligent platform iS2C2: expert assessment of the interpretability axis

To evaluate the interpretability axis of robustness for the iS2C2 platform, we assessed whether iS2C2 enhances interpretability in automated hypothesis generation through a structured, expert-based assessment[24,44,45] (Supplementary Note 1F). Fifteen experts,

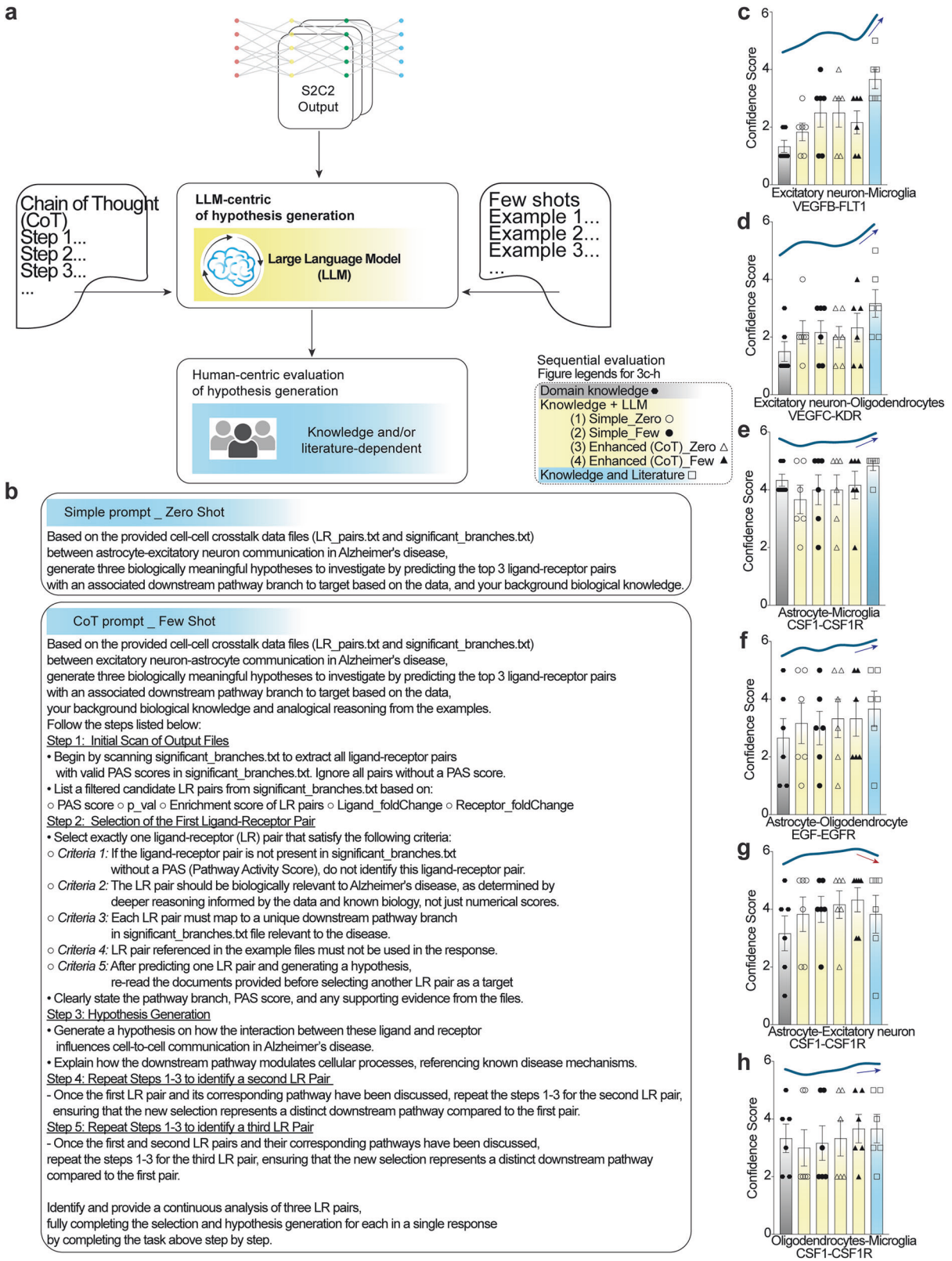


Fig. 3 Evaluation of a cointelligent platform integrating S2C2 outputs with LLM-driven hypothesis generation. **a** Schematic overview illustrating the integration of structured outputs from S2C2 with automated LLM-based reasoning and complementary manual expert evaluation. **b** Prompting strategies applied to the LLM, including simple zero-shot prompting and chain-of-thought (CoT) few-shot prompting. **c**–**h** Human-centric evaluation results for predicted ligand–receptor interactions across six representative cell–cell communication contexts. Confidence scores were assigned by domain experts under varying levels of external support (domain knowledge alone, LLM-generated hypotheses, and literature evidence)

including both male and female participants aged 25+ to 40+ years, from the Houston Methodist Hospital, Baylor College of Medicine, MD Anderson Cancer Center, and Ben Taub Hospital participated, representing diverse expertise in clinical AD neurology, neuroinformatics, computational biology, and preclinical and wet-lab research (Supplementary Fig. 17a, b). Evaluations were completed over ~1 day to 1 week for each participant.

Experts scored each prediction based on the following scenarios (Supplementary Fig. 17c): S1, evaluation based solely on domain knowledge without consulting external sources; S2, domain knowledge supplemented by LLM-generated hypotheses linking S2C2 predictions to disease-relevant information; and S3, domain knowledge combined with independently retrieved literature. Predictions were rated on a 1–5 scale for disease relevance (1 = strong disagreement, 5 = strong agreement) and assigned a consensus confidence score from 1–3 (low to high confidence) (Supplementary Fig. 17d, e).

To systematically evaluate iS2C2 and compare its performance with established frameworks, we divided the experts into three groups (Supplementary Fig. 18a) and designed four evaluation schemes for both internal comparisons of iS2C2 and for cross-framework comparisons against other baseline methods. Group 1 evaluated S1→S2→S3 sequentially; Group 2 evaluated S2→S3; and Group 3 evaluated S3→S2. This design enabled both sequential and independent assessments of iS2C2 predictions. For cross-framework comparison against S2C2, CellChat, and NicheNet, we implemented a crossover design to counterbalance scenario order effects (Task Design 3; Supplementary Fig. 18a). All experts also provided an overall consensus score across predictions (Task Design 4). This multilayered experimental framework enabled a robust and unbiased comparison of iS2C2 with existing approaches.

In Task Design 1, we use Group 1 evaluators to emulate the typical real-world process of hypothesis generation in data-driven research. Researchers often begin by applying a crosstalk algorithm to generate a large list of candidate interactions, filter these based on personal expertise (Scenario 1), and then narrow the list further following a literature review (Scenario 3). The first filtering stage is inherently biased by individual knowledge, potentially excluding valuable candidates recoverable in later stages. We tested whether LLM-assisted prompting (Scenario 2) could mitigate this bias. Across prompting strategies (simple vs. enhanced; zero-shot vs. few-shot), LLM integration increased confidence scores relative to the knowledge-only condition, with a consistent upward trajectory across settings (Fig. 3c–h, yellow-highlighted bars).

In addition, when experts consulted the associated literature, 75% (6/8) of examined *L–R* pairs reached a confidence score ≥ 3 (Supplementary Fig. 18b, purple highlighted), and all *L–R* pairs showed improvement over the knowledge-only condition (3/8 scored between 1 and 2 and 2/8 between 2 and 3; Fig. 3c–h and Supplementary Fig. 18b, blue highlighted). For instance, the VEGFB–FLT1 pair in excitatory neuron–microglia crosstalk rose from 1.33 (Scenario 1) to 2.18 with LLM-based hypotheses and to 3.67 after literature review (Supplementary Fig. 18b, arrow), illustrating how LLM-supported insights can highlight promising *L–R* pairs that might otherwise be overlooked. This demonstrates that iS2C2 can leverage LLMs (Scenario 2) to augment researchers' domain knowledge (Scenario 1) to yield results aligned with ground-truth validation (Scenario 3). An exception was the CSF1–CSF1R interaction from astrocytes to excitatory neurons—uniquely detected by iS2C2 but not by other algorithms—which received a lower score after literature review (Fig. 3g). We speculate that this reflects the scarcity of reports documenting this interaction specifically in astrocyte–excitatory neuron communication. This suggests that while the interaction itself is established, its occurrence in this particular cellular context may represent a previously uncharacterized mode of intercellular communication.

In Task Design 2, we evaluated the impact of incorporating LLM-based prompt engineering (Scenario 2) compared with domain alone (Scenario 1) or knowledge plus literature (Scenario 3) using an independent-scenario design. Specifically, as Group 1 followed S1→S2→S3, Group 2 followed S2→S3, and Group 3 followed S3→S2, by comparing the results from each scenario as the first-encountered condition (Supplementary Fig. 18a), we were able to assess the effect of each approach independently, without prior influence from the other scenarios. Consistent with Task Design 1, Scenario 1 produced lower confidence scores than Scenarios 2 or 3 (Supplementary Figs. 18c, d and 19a), reflecting the lack of LLM or literature support. Notably, incorporating chain-of-thought (CoT) reasoning in iS2C2 further increased confidence scores (Supplementary Fig. 18d). Across both sequential (Design 1) and independent (Design 2) designs, iS2C2 interpretability was robust to scenario order.

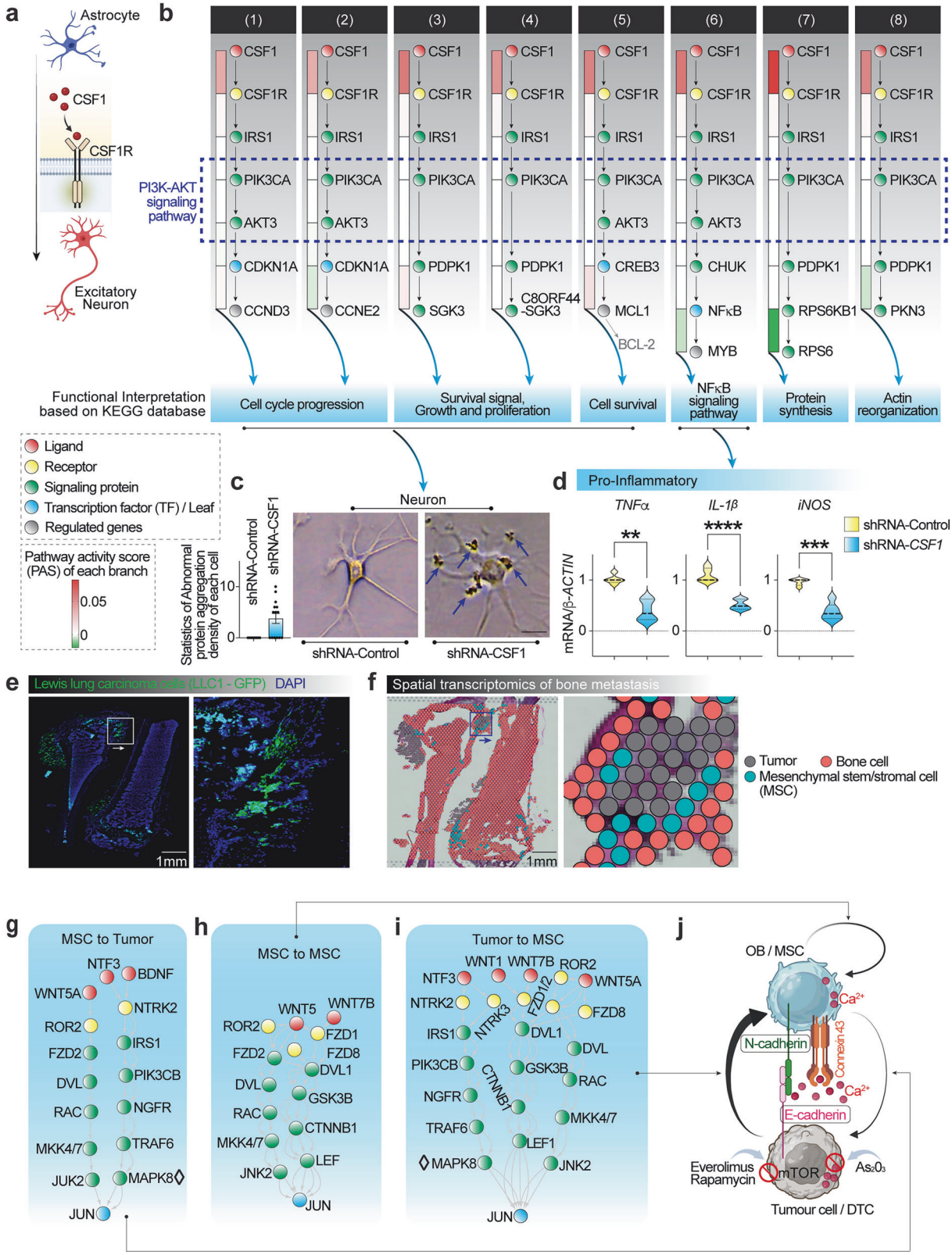
In Task Design 3, we compared the interpretability of iS2C2 with CellChat+LLM and NicheNet+LLM using evaluations from Groups 2 and 3 (Supplementary Fig. 19b). This approach minimized potential bias arising from the order of scenario presentation. In Group 2's Scenario 2, iS2C2 achieved the highest average confidence score and top average ranking and received the most first-place votes. For Scenario 3 in Group 2, NicheNet achieved the top average confidence score, while both NicheNet and iS2C2 shared the leading position in average ranking, and iS2C2 received the most first-place votes. In Group 3, iS2C2 consistently ranked first in terms of average confidence score, tied with NicheNet for the top average ranking in Scenario 3, and again secured the most first-place votes in both Scenario 2 and Scenario 3. Across both orders (Group 2: S2→S3, Group 3: S3→S2), iS2C2 repeatedly demonstrated high interpretability, frequently ranking at or near the top, and garnered the largest number of first-place votes (Supplementary Fig. 19b, arrows 1–12).

In Task Design 4, we stratified the nine evaluators from Groups 2 and 3 by their consensus scores rather than the order in which they evaluated each scenario (Supplementary Fig. 19c). Each expert assigned an overall consensus score from 1 (lowest confidence) to 3 (highest confidence) to reflect their certainty during the evaluation. The results showed that, across all consensus score groups in Scenario 2, iS2C2 achieved the highest confidence score, average ranking, and number of first-place votes. In Scenario 3, the top-ranked framework varied by consensus group: CellChat ranked first in the consensus score 1 group (1 evaluator), NicheNet ranked first in the consensus score 2 group (3 evaluators), and iS2C2 ranked first in the consensus score 3 group (5 evaluators). Overall, iS2C2 maintained high average confidence scores across all consensus strata in Scenario 2 (Supplementary Fig. 19c, arrows 13–18) and strong performance in Scenario 3 (Supplementary Fig. 19c, arrows 19–24). Notably, more experts with high consensus scores assigned their top ratings to iS2C2 (Supplementary Fig. 19c, arrows 25–26).

In summary, expert-based assessment found that iS2C2 consistently achieved high interpretability, one of the key metrics of robustness, along with confidence scores across a range of evaluation designs and scenarios. Performance remained stable regardless of scenario order or assessment approach, indicating that iS2C2 is a robust, practical, and reliable platform for automated hypothesis generation in biomedical research.

Validation of iS2C2–generated mechanism hypotheses for Alzheimer's disease

In the first case study, we performed a secondary analysis of AD snRNA-seq data[35,46,47] covering distinct regions of the human brain to evaluate data-driven organ zonation using S2C2 (Supplementary Fig. 20). By integrating datasets that capture the spatiotemporal details of AD progression, S2C2 identified the top cell–cell crosstalk interactions among distinct cell types and enriched downstream pathways (Supplementary Fig. 21). Notably,



the CSF1–CSF1R interaction emerged as the top *L–R* interaction, with marked downstream branching between astrocytes and excitatory neurons in the prefrontal cortex (Supplementary Figs. 20, 22 and 23)—an interaction typically associated with microglia.[48]

A stepwise analysis of Tsai's data[35] identified 119 *L–R* pairs (Supplementary Fig. 24a), with 35 showing statistically significant downstream pathway activation (P value < 0.05) (Supplementary Fig. 24b). Among these, CSF1–CSF1R demonstrated one of the highest fold-changes (Supplementary Fig. 24b) and was involved

Fig. 4 Application of iS2C2 in dissecting molecular mechanisms in Alzheimer's disease and metastatic bone cancer. **a** Crosstalk between astrocytes and neurons. A schematic diagram illustrating the predicted crosstalk between astrocytes and neurons via the CSF1–CSF1R signaling pathway as identified by S2C2. **b** Predicted downstream pathway branches of CSF1–CSF1R. The results from S2C2 predict the downstream branches of the CSF1–CSF1R pathway, including the PI3K–AKT signaling between astrocytes and neurons. Each branch is annotated according to the KEGG database. **c** Neuronal response to CSF1 knockdown. Representative images and statistical data showing neurons with abnormal protein aggregation following shRNA-CSF1 treatment (right) compared to healthy controls (left). CSF1 in astrocytes was knocked down using shRNA-CSF1, and the conditioned medium from these astrocytes with reduced CSF1 secretion was used for neuron culture. This led to a decrease in the CSF1–CSF1R downstream signaling levels due to the reduced interaction between CSF1 and CSF1R. Scale bar = 10 μm. **d** RT-PCR analysis of proinflammatory gene expression. Summary of three to six independent RT-PCR studies assessing the expression of proinflammatory genes, including *TNF-α*, *IL-1β*, and *iNOS*, with or without shRNA-CSF1 treatment. **e** Confocal imaging of bone metastasis. Representative confocal image of Lewis lung carcinoma (LLC1-luciferase-GFP)-mediated mouse bone metastasis with DAPI nuclear staining. Whole tibia and femur image (left), scale bar = 1 mm; partial enlargement (right). **f** Spatial map of cell types in bone metastasis. Representative spatial map depicting different cell types in bone metastasis. Cells are colored by their cluster identities. Whole tibia and femur image (left), the same tissues with (**e**) scale bar = 1 mm; partial enlargement (right) showing the crosstalk between tumor cells and mesenchymal stem/stromal cells (MSCs). **g–i** Predicted pathway branches with JUN as a transcription factor in bone metastasis. The results from S2C2 predicted pathway branches with JUN as the transcription factor in the context of bone metastasis. Branches from MSCs to tumor cells (**g**) from MSCs to MSCs (**h**) and from tumor cells to MSCs (**i**). **j** Crosstalk between MSCs and tumor cells. A schematic diagram illustrating the predicted crosstalk between MSCs and tumor cells

in multiple downstream pathways (Supplementary Fig. 25), including eight branches associated with the PI3K–AKT signaling pathway (Supplementary Table 3; Fig. 4a, b). By directly integrating the S2C2-derived crosstalk results from Tsai's dataset into our cointelligent platform, we identified CSF1–CSF1R, mediated by the PI3K–AKT pathway, as the top-ranked interaction (Supplementary Fig. 26).

To validate these predictions, we performed knockdown experiments targeting CSF1 in astrocytes and/or CSF1R in neuronal cell lines (Supplementary Fig. 27). Predicted downstream branches revealed associations with KEGG functions, including cell cycle progression, cell growth/proliferation related to neuronal protection, and NFκB-mediated proinflammatory responses (Fig. 4b). Knocking down CSF1 in astrocytes severely impaired CSF1–CSF1R-mediated neuronal protection function, leading to abnormal protein aggregation and subsequent apoptosis in neurons (Fig. 4c). Additionally, blocking the CSF1–CSF1R interaction diminished the PI3K–AKT-induced proinflammatory response (Fig. 4d), consistent with previously reported proinflammatory signatures in neurons.[49] These results confirm that the downstream pathway predictions made by S2C2 are experimentally verifiable and highlight the potential of our cointelligent platform to identify novel cell–cell interactions. Importantly, this interaction pattern suggests that divergent cell types can share convergent crosstalk pathways. For instance, both astrocytes and microglia exhibit the CSF1–CSF1R interaction, extending potential therapeutic targets beyond microglia to include astrocytes and providing evidence that S2C2-predicted crosstalk patterns can have implications for drug-target selection and treatment strategies.

Validation of iS2C2–generated mechanism hypotheses for bone metastasis

In the second case study, we investigated the co-migration of disseminated tumor cells (DTCs) and mesenchymal stem cells (MSCs) in the metastatic progression of breast cancer to bone marrow, with their linkage and colocalization facilitated by E-cadherin and N-cadherin.[50,51] While the promotion of N-cadherin expression in MSCs to enable their epithelial-to-mesenchymal transition (EMT) remains unclear, we applied S2C2 to our integrated multimodality dataset from a bone metastasis model induced by Lewis lung carcinoma cells (LLC1-GFP) (Fig. 4e). This dataset included mouse ST (Fig. 4f), scRNA-seq (Supplementary Fig. 28a), and laser-capture microdissection followed by transcriptome profiling (LCM-seq), enabling precise localization of metastatic tumors with their spatial context.

Using S2C2, we first identified clusters containing DTCs and MSCs based on ST data, revealing pronounced regulation by the CTNNB1 and JUN transcription factors (Supplementary Fig. 28b), both crucial for cadherin regulation.[52] Further scRNA-seq analysis distinguished adjacent DTCs and MSCs (Supplementary Figs. 28a and 29) and revealed that JUN-mediated crosstalk from tumors to MSCs was more prevalent than that from MSCs to tumors or within MSCs (Fig. 4g–i). This suggests that the high expression of N-cadherin in MSCs is induced by tumor-MSC communication initiated by DTCs (Fig. 4j).

These findings validate that S2C2 can effectively harness spatial and single-cell transcriptomics to elucidate the molecular mechanisms underlying observed phenotypic changes. Simultaneously, when we streamlined the complex crosstalk between MSCs and tumor cells and forwarded the S2C2 prediction to our cointelligent platform, we discovered that the upregulation of JUN is not solely induced by WNT family signaling; the NTF3-mediated pathway also contributes to this process (Supplementary Fig. 30a). This suggests that pathways mediated by the NTF family may promote tumor metastasis. Recent studies have detailed the role of the NTF family in enhancing EMT, including dynamic cadherin regulation,[53–55] which aligns with our findings.[50,51] The pathway patterns revealed by the cointelligent platform further indicate that designing therapeutic interventions to inhibit tumor metastasis may require blocking multiple inducing factors or ligands.

To simultaneously inhibit the NTF3 and WNT pathways, we utilized the drug-prediction module in S2C2 (Supplementary Fig. 3). Tamoxifen was identified as a candidate, as it can regulate mitogen-activated protein kinase 8 (MAPK8) within the NTF3 pathway (Fig. 4i and Supplementary Fig. 30b) and serves as a direct inhibitor of estrogen receptor alpha (ERα) (Supplementary Fig. 30b). ERα is known to regulate *WISP2* (WNT1-inducible signaling pathway protein 2), a downstream effector of the WNT pathway,[56,57] suggesting that tamoxifen could modulate both NTF3- and WNT-driven signaling.

To validate this hypothesis, ERα-positive cells were treated with tamoxifen. The expression of *PGR*, a direct downstream target of ERα, was markedly reduced (Supplementary Fig. 30c). Notably, *WISP2* expression was also decreased (Supplementary Fig. 30d). Building on these in vitro findings, we assessed tamoxifen in the LLC1 bone-metastasis mouse model (Fig. 4e–j). The in vivo results indicated that tamoxifen has the potential to inhibit rapid bone metastatic progression at an exceptionally early stage in the LLC1 model (Supplementary Fig. 30e). These findings validate the NTF and WNT pathway patterns predicted by iS2C2 in the context of bone metastasis and highlight the platform's utility in uncovering actionable mechanistic insights for therapeutic intervention.

DISCUSSION

The rapid advancement of generative artificial intelligence (AI) offers unprecedented opportunities to decipher complex disease mechanisms and accelerate therapeutic innovation through the

integration of multimodal data, advanced computational models, and deep domain knowledge.[37,58–61] Despite these transformative possibilities, existing methodologies often lack cohesion across individual algorithmic components, limiting their reliability and interpretability. Chaining disparate algorithms—each optimized for an isolated task—often fails due to inherent mathematical complexity and context dependency in biomedical data, challenges that current LLMs struggle to reliably overcome.[62–64] Addressing these critical limitations requires an intentionally designed AI architecture that ensures coherent integration and effective communication among its AI components.

Our co-intelligent platform, iS2C2, embodies this principle by tightly coupling a mechanistically grounded cellular crosstalk model (S2C2) with advanced LLM reasoning capabilities via prompt engineering. This integration improves the accuracy, interpretability, and efficiency of generating mechanistic hypotheses from single-cell RNA-seq and spatial transcriptomics data, enabling systematic identification and prioritization of therapeutic targets. In the meantime, additional challenges arise from the integration and behavior of LLMs. One key concern is hallucinations—LLMs may generate outputs not grounded in the input data or known biology.[25,26,65,66] For example, we observed that when the underlying algorithms did not achieve perfect accuracy, the LLM predicted *L–R* pairs that were not present in the original crosstalk analysis. While support from the literature and expert review is valuable, we emphasize that experimental validation remains the ultimate standard for ensuring biological plausibility. To further reduce the risk of hallucinations, we recommend that all LLM-generated results undergo manual review before proceeding to experimental validation.

One way to address this challenge is to employ domain-specific visualizations, such as heatmaps of *L–R* expression and network diagrams of downstream pathway alterations,[12,16] which can enhance the effectiveness and interpretability of the results. iS2C2 addresses these challenges by combining S2C2 with prompt engineering. S2C2 enhances crosstalk prediction by defining pathway branches and integrating *L–R* expression with alterations in downstream signaling pathways. This approach enables the identification of curated *L–R* interactions, prioritization of downstream effectors, and inference of signaling modules. Building on S2C2, iS2C2 incorporates prompt engineering specifically to improve the accuracy, reproducibility, and interpretability of LLM-based predictions. This seamless integration enables automated generation of biologically grounded and experimentally testable hypotheses, substantially accelerating translational biomedical discovery. The applications of iS2C2 extend from hypothesis generation and pathway discovery to the development of targeted intervention strategies and the mechanistic dissection of pathological processes using both single-cell RNA-seq and spatial transcriptomics data.

Validation studies clearly illustrate the novelty and translational impact of iS2C2. In AD, the platform identified an unprecedented astrocyte-to-neuron signaling axis mediated by CSF1 and CSF1R. Traditionally, research has primarily focused on the role of CSF1R in microglia,[67–70] including its involvement in microglial proliferation, nonplaque-associated microglia (PAM) to PAM transition, transgenic depletion models, and microglial homeostasis, with limited attention given to astrocytic sources of CSF1. Neurons and astrocytes have both been proposed as potential sources of CSF1.[71,72] S2C2's prediction that astrocyte-secreted CSF1 activates neuronal CSF1R, triggering downstream PI3K–AKT signaling. This is consistent with established CSF1R pathways in microglia.[73,74] Although this crosstalk received top-ranked scores in both S2C2 predictions and LLM-based hypothesis generation, expert literature review revealed minimal prior documentation (Fig. 3g). This novel discovery highlights iS2C2's capability to uncover overlooked, cell-specific signaling mechanisms and promotes future experimental validation. Moreover, we are extending this approach to investigate multicellular cascading crosstalk, as exemplified by a validated case involving microglia–astrocyte–excitatory neuron interactions (Supplementary Fig. 31), supported by previous studies.[75,76]

In the context of bone metastasis, iS2C2 identified cadherin-mediated communication between DTCs and MSCs,[50,51] which is jointly regulated by WNT and NTF signaling through downstream activation of the transcription factor JUN (Fig. 4g–j). This convergence of pathway signals (Fig. 2f and Supplementary Fig. 28a) suggests that effective early intervention in bone metastasis may require simultaneous inhibition of both the WNT and NTF pathways. Supporting this, recent studies have shown that targeting the NTF family can effectively suppress bone metastasis.[53–55] Our experimental validation further revealed that tamoxifen, traditionally known for ERα blockade, simultaneously modulates the NTF and WNT pathways via novel targets MAPK8 and WISP2 (Supplementary Fig. 30). This previously unrecognized dual-pathway modulation is consistent with clinical evidence that tamoxifen reduces the risk of metastatic disease.[77] This early intervention led to a marked reduction in bone metastatic progression in the LLC1 model. These findings highlight the value of iS2C2 in identifying combinatorial therapeutic strategies targeting multiple signaling axes. By revealing complex pathway interdependencies, our approach offers a powerful platform for effective drug intervention.

Despite its demonstrated utility, we still recommend that all iS2C2-generated results undergo manual review prior to experimental validation to minimize the risk of hallucinations. Additionally, we advise running iS2C2 multiple times and selecting consistently repeated predictions to ensure high reproducibility before proceeding to functional validation. One factor contributing to output variability is the temperature parameter, which governs the randomness of LLM-generated responses. In iS2C2, this parameter is user-adjustable for all supported LLM backends, enabling users to reduce or increase stochasticity as needed. Temperature tuning, therefore, provides a practical option for managing variability, although some degree of randomness is unavoidable in current LLM architectures. Another constraint lies in LLM bias: general-purpose models may reflect imbalances in their training data and lack precision in biomedical contexts.[22,25,78,79] These issues are further compounded by the absence of domain-specific fine-tuning. To mitigate these risks, we suggest prompt engineering, as more specific prompts and step-by-step instructions can explicitly encourage neutrality or consideration of multiple perspectives to help counteract certain biases.[39,80,81]

Moreover, we suggest constructing a PubMed-derived knowledge graph to guide LLM reasoning, as this can further help address hallucinations and biases.[82–85] By connecting LLM responses to curated and structured information from PubMed, model outputs are more likely to remain consistent with established biomedical knowledge, and factual errors can be minimized since the knowledge graph acts as an external fact-checking resource. If the knowledge graph is built from a diverse and comprehensive set of PubMed articles, it can help balance some biases inherent in the training data by anchoring reasoning to the broader scientific consensus. Additionally, such an approach increases transparency, as model outputs can be traced back to specific nodes or edges in the knowledge graph, supporting interpretability and validation.

Importantly, we have observed continuous improvement in prediction accuracy and reproducibility with newer LLM models, highlighting ongoing advancements in AI technology (Supplementary Fig. 15a, b). Nonetheless, current LLMs still lack transparency, which limits interpretability.[25,86–88] Moving forward, integrating explainable AI frameworks and domain-specific fine-tuning will further bolster transparency, robustness, and trustworthiness in biomedical hypothesis generation.[89–92]

In summary, by systematically integrating robust crosstalk mechanistic modeling with advanced interpretative reasoning, the cointelligent iS2C2 platform provides a foundation for a new class of AI-augmented biological discovery tools. These tools promise to democratize access to complex data insights, accelerate hypothesis validation cycles, and unlock new therapeutic opportunities across diverse disease areas such as neurodegeneration, cancer, and beyond.

MATERIALS AND METHODS

S2C2 database retrieval and processing

The *L–R* database was developed following the approach outlined in our previous study,[13] with updated information incorporated from KEGG (updated as of Jan 2024). Signaling and gene interaction data were sourced from KEGG and Ingenuity Pathway Analysis (IPA, version 22.0.2). Self-binding interactions were excluded from the integrated network to create simplified, unidirectional pathway networks. Pathways with identical or similar names from multiple sources were retained and labeled according to their originating database. These combined resources were then merged into a comprehensive gene signaling database to support S2C2 modeling. Each pathway in the S2C2 database was transformed into a graph, $G = (V, E, W)$, where the node set V represents the genes in the pathway, the edge set E denotes the interactions among these nodes, and W indicates the context-specific weight of each edge based on the nature of the interaction as defined by the gene signaling database. Positive interactions, such as activation, causation, binding, phosphorylation, and expression, were assigned a positive weight (+1). In contrast, inhibitory interactions, including inhibition, deactivation, dephosphorylation, and dissociation, were assigned a negative weight (−1). For interactions categorized as indirect or unspecified, a neutral weight (+0.5) was applied.

Preprocessing AD single-nucleus RNA sequencing (snRNA-seq) data from Tsai's dataset[35]

The raw cell counts were obtained from Synapse (syn18485175) and preprocessed according to the protocols outlined in the original paper, utilizing the Seurat package.[93] Differentially expressed genes (DEGs) were identified using the FindMarkers function in Seurat. Four brain cell types were annotated: astrocytes, microglia, inhibitory neurons, and excitatory neurons.

Preprocessing AD snRNA-seq data from Swarup's dataset[46]

The dataset was accessed via Synapse (syn22079621), and the preprocessing steps provided by the original authors were followed using code from their GitHub repository (https://github.com/swaruplab/Single-nuclei-epigenomic-and-transcriptomic-landscape-in-Alzheimer-disease/), resulting in a fully processed Seurat object. For comparison, DEGs between the control group and AD patients were identified for key cell types (astrocytes, microglia, inhibitory neurons, and excitatory neurons) using the FindMarkers function in Seurat.

Preprocessing AD snRNA-seq data from the Kampmann dataset[47]

Data were retrieved from Synapse (syn21788402), focusing on two brain regions: the superior frontal gyrus (SFG) and the entorhinal cortex (EC). Each region was treated as a separate dataset, labeled Leng_EC and Leng_SFG. All the data were processed using the code provided by the original authors, generating two distinct Seurat objects. AD pathology was classified based on Braak staging, where a Braak stage of zero indicates early AD pathology and a stage of six indicates moderate to severe pathology. For comparative analysis, DEGs between Braak stages zero and six were identified using Seurat's FindMarkers function.

Preprocessing of TICAtlas single-cell RNA sequencing (scRNA-seq) data

Processed RDS files for the TICAtlas were obtained from Zenodo (https://doi.org/10.5281/zenodo.4263972) as described by Nieto et al.[36] For this study, we specifically extracted and analyzed data from breast cancer and NSCLC cohorts. Analysis focused on characterizing cellular crosstalk within the immune microenvironment; specifically, we investigated interactions between M2 tumor-associated macrophages (M2 TAMs) and natural killer (NK) cells in breast cancer and interactions between M2 TAMs and conventional dendritic cells (cDCs) in NSCLC.

Single-cell RNA-seq data analysis of bone metastasis

Bone marrow single-cell gene expression data were processed using Cell Ranger 7.1.0 software, with alignment to the mouse genome reference, refdata-gex-mm10-2020-A. Quality control was conducted via the Seurat pipeline, excluding cells with more than 2500 or fewer than 200 unique features, as well as those with over 5% mitochondrial content. Data normalization was performed using the global-scaling LogNormalize method, which adjusts the expression levels of each feature by dividing by the total expression for each cell, multiplying by a scale factor (default 10,000), and applying a log-transformation. A linear transformation (scaling) was subsequently applied, followed by dimensionality reduction using principal component analysis (PCA). Cell clustering was achieved through Seurat's FindClusters function, utilizing a graph-based approach to group cells. Nonlinear dimensional reduction was conducted with UMAP to visualize clusters. DEGs for each cluster were identified using the FindMarkers function, and cell types were annotated based on known markers: erythroid cells (Tfrc, Gypa, Gata1, Klf1, Hbb-bs, Hba-a1/a2), MSCs (Nt5e, Thy1, Ly6a), T cells (Cd4, Cd3e), B cells (Cd19), endothelial cells (Cd34, Tek, Kdr), myeloid cells (Cd33, Itgam, Ly6g), and macrophages (Adgre1, Csf1r, Cd68).

Spatial transcriptome data analysis of bone metastasis

Spatial transcriptomics data from the early bone-metastasis mouse model induced by Lewis carcinoma cells (LLC1-GFP) were generated using the 10x Visium platform. FASTQ files from four samples were aligned to the mouse refdata-gex-mm10-2020-A reference genome using Space Ranger v2.1.0. Initial preprocessing of the spot-by-gene expression data was performed with Seurat, and normalization was applied to account for variance in sequencing depth across data points. The four samples were then integrated into a single dataset using Seurat's FindIntegrationAnchors function, where pairs of similar cells (anchors) were identified across datasets through reciprocal PCA and mutual nearest neighbors methodologies. These anchors enabled the alignment and integration of the datasets, mitigating batch effects while preserving shared biological structures. Dimensionality reduction and clustering were carried out using the same approaches as described for single-cell data analysis.

Multimodality data integration of bone metastasis

To identify tumor regions in the spatial transcriptomic tissue slices, immunofluorescence (IF) images with GFP signals (LLC1-GFP), marking tumor cells, were generated from tissue slices adjacent to those used for spatial transcriptomics. Tumor-enriched regions in the IF images were aligned with the spatial transcriptomics (ST) images using ImageJ. To locate stroma- and immune-enriched areas, scRNA-seq data were integrated with the ST data using the anchor-based integration workflow from Seurat v3, enabling probabilistic transfer of annotations from a reference dataset to a query set. This process facilitated a probabilistic classification of cell types for each ST spot according to scRNA-seq-derived cell type identities.

Spatial measurement of cell–cell distance/clustering in ST data using Ripley's K function

Analyzing all pairwise cell–cell communications demands substantial computational time and resources, especially for spatial transcriptomic data where dozens of cell groups may be identified on each tissue slide, potentially resulting in hundreds of cell–cell interaction pairs to predict cell communication. However, autocrine and paracrine communications typically occur among cells within approximately 200 µm.[94] Thus, running crosstalk algorithms on all possible cell pairs is unnecessary and inefficient. To address this, a bivariate K function was introduced to assess the spatial relationships between stromal and tumor cell groups before applying S2C2 to spatial transcriptomics data. The K function is defined as:

$$K_{i,j}(r) = \frac{1}{D_i D_j A} \sum_k \sum_l I\left(d_{i_k,j_l} < r\right)$$

where:

- $I[\bullet]$ denotes the indicator function, which equals 1 if the condition is true and 0 otherwise.
- D_i and D_j represent the densities of cell group i and cell group j in each ST sample.
- d_{i_k,j_l} is the distance between the kth cell of group i and the lth cell of group j.
- A refers to the area of the tissue region in each ST sample.
- r is the selected radius (default is 200 µm), representing the potential *L–R* interaction range for local paracrine pathways when calculating the bivariate *K* function.

A higher K score indicates greater clustering of the two groups of cells within the specified radius in each sample. To assess whether the two cell types are significantly clustered across all samples, a normalized K score is calculated for each sample and subsequently averaged over all ST samples. Monte Carlo simulations are performed to determine the upper and lower bounds of the 95% confidence interval for the K values. The normalized *K* score is then defined as: $K_{i,j}^{norm} = \frac{2K_{i,j} - K_{i,j}^{upper} - K_{i,j}^{lower}}{K_{i,j}^{upper} - K_{i,j}^{lower}}$

where:

- $K_{i,j}^{upper}$ and $K_{i,j}^{lower}$ are the upper and lower bounds of the 95% confidence interval obtained from the Monte Carlo simulations.
- $K_{i,j}$ is the observed K score between cell groups i and j.

This normalization enables direct comparisons across multiple samples by ensuring a consistent range of variation in spatial clustering. A normalized K score greater than 1 indicates significant clustering between groups i and j, while a score less than -1 indicates significant dispersion between the groups. Scores between -1 and 1 suggest nonsignificance. Only cell groups with an average normalized K score greater than 1 across all samples are recommended for spatial crosstalk analysis.

Preparation of S2C2 input

A standard Seurat object in R, containing gene expression data in the assay slot and cell type annotations in the metadata slot, was used as input for S2C2. This object is generated through conventional workflows applicable to both single-cell and spatial transcriptomics data, including data normalization, PCA-based dimensionality reduction, clustering, and annotation. To compare different conditions, sample condition information, such as AD vs. Control in the AD crosstalk use case, must be included in the metadata slot. If this information is absent, comparisons are performed within the same condition, as shown in the bone metastasis use case. For both the Java-based GUI and the CLI, an RDS file saved using the Seurat saveRDS function can be used as input.

Prediction of ligand–receptor (*L–R*) interactions with S2C2

L–R interactions between sender and receiver cells were determined based on their expression patterns, characterized by fold change (FC) and the percentage of expressed cells (PE) in each cell group. Ligands were filtered using default thresholds: an average log2-fold change greater than 0.2, expression in more than 0.5% of the sender cell population, and a p value below 0.05. Receptors were selected if they were expressed in more than 0.5% of the receiver cell group, as receptor upregulation is not required to activate downstream pathways. An enrichment score (ES) was calculated for each *L–R* pair to further refine the interactions. The ES is defined as:

$$ES(C_s, C_r, L, R) = mean[PE_L^{C_s} \times \log(FC_L^{C_s}) + PE_R^{C_r} \times log(FC_R^{C_r})]$$

where:

- $FC_L^{C_s}$ and $PE_L^{C_s}$ denote the average fold change and the percentage of expressed cells for ligand L in the specified sender cell C_s.
- $FC_R^{C_r}$g and $PE_R^{C_r}$ denote the corresponding values for receptor R in receiver cell C_r.

When different sample conditions are provided, the fold change is calculated by comparing the mean gene expression of the cell groups across conditions (e.g., AD vs. control). If no sample conditions are specified, the fold change is calculated by comparing the value across different cell groups (e.g., tumor vs. nontumor). The enrichment score represents the strength of the interaction between the ligand and receptor. A higher enrichment score indicates that both ligand and receptor exhibit larger fold changes and are expressed in more cells. Only *L–R* pairs with a positive enrichment score were predicted as potential intercellular communication candidates and exported to the "LR_pairs.txt" output file. Additionally, the significance of the enrichment score for the (L, R) pair from sender cell group C_s to receiver cell group C_r was calculated to measure the overall contribution of the (C_s, C_r) cell pairs to the (L, R) interaction among all possible cell pairs. This significance score is defined as:

$$ES_{sig}(C_s, C_r, L, R) = \frac{\sum_{i,j} I[ES(C_i, C_j, L, R) \geq ES(C_s, C_r, L, R)]}{\sum_{i,j} 1} \in (0, 1]$$

where:

- $I[\bullet]$ is the indicator function.
- $i, j \in \{1, 2, \ldots, k\}$ represents the set of all k cell types in the dataset.

A lower value of ES_{sig} indicates a more significant contribution of the cell type pair (C_s, C_r) to the (L, R) interaction among all possible cell pairs. This specificity score is included in the "LR_pairs.txt" output file for reference.

Prediction of activated pathway branches with S2C2

Tens of *L–R* interactions are typically included within each established pathway in KEGG or IPA, followed by various signal flow paths triggered by each *L–R* interaction, potentially involving hundreds of genes contributing to each pathway. Therefore, a single activation score for each pathway is not considered insightful, as different activation patterns regulating various biological functions of cells are encompassed within the same comprehensive pathway. To specify which specific downstream flows are triggered by each *L–R* interaction, the concept of

pathway branches is introduced. Each pathway's branches are defined as the subgraph of its pathway graph $G = (V, E, W)$, denoted by $P = (V_p, E_p, W_p)$,where:

- $V_p = \{v_0, v_1, \ldots, v_m\} \subseteq V$ denotes the node set of subgraph P.
- $E_p = \{e_0, e_1, \ldots, e_n\} \subseteq E$ denotes the edge set of subgraph P.
- $W_p : E_p \to \{1, -1, 0.5\} \subseteq \mathrm{W}$ denotes the weight of E_p in subgraph P.
- $\deg(\mathrm{v_i}) = 2$ for i in $\{1, 2, \ldots, \mathrm{m}-1\}$ or 1 for i in $\{0, \mathrm{m}\}$ denotes the node degree, reflecting the number of edges linked to the node.
- $\deg(\mathrm{e_j}) = 2$ for j in $\{0, 1, \ldots, \mathrm{n}\}$ denotes the edge degree, reflecting the number of nodes connected to the edge.

Here, $v_0 = L, v_1 = R$ represents the ligand and receptor pair with an enrichment score greater than 0, while v_m represents the transcription factor or leaf node, which has no outgoing edges as defined in KEGG and IPA. The set of $\{v_2, v_3, \ldots, v_{m-1}\}$ denotes all signaling genes linking the (L, R) pair to transcription factors. Note that $n = m - 1$, indicating the property of a graph path. The pathway branches are generated by linking (L, R) to each transcription factor using the shortest_paths function in igraph (R package version 2.0.3, https://CRAN.R-project.org/package=igraph). Spearman correlation scores for genes along the pathway branch were calculated to assess the direct correlation between consecutive genes on each edge of the pathway branch. Finally, a PAS for pathway branch P triggered by interaction of (L, R) is calculated as:

$$PAS(L,R,P) = \lambda(logFC_L + logFC_R) \times w_{L,R} + (1-\lambda)\sum_{i=1}^{m-1} w_{e_i} \times cor(v_i, v_{i+1}) \times logFC_{v_i}$$

where:

- FC_L, FC_R and FC_{v_i} are the fold changes in ligand, receptor, and downstream branch genes v_i.
- $w_{e_i} \in \{-1, 1, 0.5\}$ is the edge weight in pathway branch P.
- $w_{L,R} = \frac{\sum_{i=1}^{m-1} w_{e_i} * cor(v_i, v_{i+1})}{n}$ denotes the weight of L–R interactions. Note that this differs from w_{e_0}, which equals 1.
- λ is a hyperparameter that controls the balance between the influence of the L–R pair and their downstream pathway branches on cell communication signaling activation. When λ is set to 0.5 (default), the L–R pair and the downstream pathway genes contribute equally to the PAS. A λ value of 0 indicates that the score is based entirely on downstream pathway genes, ignoring the L–R influence. A value of 1 makes the score depend solely on the L–R pair.

To identify significant pathway branches, a permutation test is performed by S2C2 for each pathway branch j over n iterations. In the ith iteration, random genes are sampled to replace the pathway genes, and the pathway activity score $\mathrm{PAS}_i^{(j)}$ is calculated, with the default $n = 500$ iterations. An empirical distribution of $\mathrm{PAS}^{(j)}$ is formed from these 500 scores. This process is repeated for each pathway branch, and the p value for pathway branch j is calculated by:

$$p^j = \frac{\sum_i I(\mathrm{PAS}_i^{(j)} > \mathrm{PAS}^{(j)})}{n}.$$

A smaller p value indicates higher significance in pathway alterations.

Prioritization of ligands and terminal nodes

To facilitate efficient experimental validation of cell communication predictions, the predicted ligands, pathway branches, and terminal nodes, such as transcription factors or their targets, need to be ranked. Beyond enrichment scores, ligands can also be ranked by the number of associated downstream activated pathway branches and downstream genes. Higher-ranked ligands may indicate that more signaling pathways are triggered, suggesting the regulation of various cellular biological processes. Similarly, transcription factors can be ranked, with higher-ranked factors reflecting converged signaling transductions from diverse L–R interactions.

Gene coverage enhancement to expand the 1k gene list to an 18k gene list

A random selection of 1000 genes was drawn from the original 18,000-gene list in the Tsai dataset.[35] Subsets containing these 1000 genes were generated for both the Tsai and Swarup[46] datasets. The H5ad files for these subsets, as well as the original datasets, were loaded using the Scanpy package in Python. Data normalization was performed using total count normalization, followed by log normalization. The data were then embedded using the tasks.embed_data function from scGPT.[37]

Two deep learning models, MappingNetwork and DecoderNetwork, were developed to learn from the embedding space. In MappingNetwork, the scGPT embeddings from the 1000-gene dataset were taken as input and passed through a multilayer perceptron (MLP) with four layers to predict the scGPT embeddings from the original 18,000-gene dataset. The structure for each layer $i \in \{1, 2, 3\}$ is

$$z_i = W_i h_{i-1} + b_i (Linear\ Transformation)$$

$$a_i = \sigma(z_i) (Activation\ Function)$$

$$\widetilde{a_i} = Batch\ Norm(a_i) (Batch\ Normalization)$$

$$h_i = Dropout(\widetilde{\mathrm{a}}_\mathrm{i}) (Dropout)$$

where:

- $h_0 = E_{1000}$ represents the embeddings of 1000 genes by scGPT.
- $\widetilde{E}_{18,000} = h_4 = W_4 h_3 + b_4$ represents the predicted embeddings of 18,000 genes by scGPT.

In DecoderNetwork, the output from MappingNetwork, along with the original 1000-gene expression profile, was taken as input and passed through an MLP with five layers to decode the embeddings back to the original gene expression for the 18,000 genes. The structure for each layer $i \in \{1, 2, 3, 4\}$ is

$$z_i' = W_i' h_{i-1}' + b_i' (Linear\ Transformation)$$

$$a_i' = \sigma(z_i') (Activation\ Function)$$

$$h_i' = Dropout(a_i') (Dropout)$$

where:

- $h_0' = \mathrm{Concat}(\mathrm{G}_{1000}, \widetilde{\mathrm{E}}_{18,000})$ represents the concatenated vector of the original gene expression values for 1000 genes and the predicted embeddings from the MappingNetwork.
- $\widetilde{G}_{18,000} = h_5' = W_5' h_4' + b_5'$ represents the predicted gene expression value of the 18,000 genes.

The number of layers for both the MappingNetwork and DecoderNetwork was determined based on the dimensions of the input and target outputs. The MappingNetwork consists of four layers, where the number of neurons doubles in each layer, starting from a 512-dimensional input, increasing to 1024, then 2048, before scaling back down to 1024, and finally to 512. In

contrast, the DecoderNetwork uses five layers to progressively upscale the combined input ($512 + 1000 = 1512$ dimensions), doubling the neurons in each layer and ultimately producing an 18K-dimensional output. The LeakyReLU activation function, with a slope of 0.01, was applied in the MappingNetwork to prevent neuron inactivity (dead neurons). In contrast, the simpler and more efficient ReLU activation function was used in the DecoderNetwork, as the model's overall capacity could tolerate a few inactive neurons without significantly impacting learning. BatchNorm was not applied to the DecoderNetwork due to the skewness in the target gene expression data. Both models were trained using mean squared error loss and optimized with the Adam optimizer at a learning rate of 0.001. Training was conducted on a cloud GPU server equipped with two RTX A5000 GPUs connected via NVLink. The Tsai dataset was split into 80% for training and 20% for validation, with a batch size of 64. The Swarup dataset was used for independent testing. Both models were trained for 50 epochs.

Model parameter sensitivity analysis

The hyperparameters of S2C2 were tested to evaluate how the predicted *L–R* interactions and activated pathway branches were affected. λ values ranging from 0 to 1 were tested in increments of 0.05, with the number of downstream pathway branches being used as the performance metric. Log-fold change thresholds between 0.10 and 0.50 were tested in increments of 0.05. As the log-fold change threshold was increased, both the number of *L–R* pairs and the number of significant pathway branches decreased. Percent expression threshold values ranging from 0.5 to 5% were tested in increments of 0.5%. When higher percent expression thresholds were applied, a decrease in both the number of *L–R* pairs and the significant pathway branches was observed.

In silico prediction and validation with pathway perturbation using S2C2. In silico experiments were conducted to evaluate the reliability of S2C2 in pathway prediction by manipulating the expression levels of *L–R* pairs and downstream genes (Supplementary Figs. 11 and 12). For the MAPK and NFκB signaling pathways, which were not identified in the original S2C2 predictions, a manual amplification of a log(1.5) increase in gene expression was applied to the ligands and receptors in these pathways, respectively. Conversely, for the PI3K–Akt and RAS pathways, which had been identified by S2C2, manual reductions were introduced to the ligands and receptors. S2C2 was then applied to the modified datasets, and the number of identified *L–R* pairs, significant downstream branches, unique genes within those branches, and PASs were compared with the original predictions. Furthermore, 15 downstream genes in the PI3K–AKT pathway, excluding ligands and receptors, were inhibited, and similar comparative analyses were performed to assess whether S2C2 is sensitive to downstream gene changes.

Following each perturbation, the modified expression matrix was processed using S2C2. The outputs were compared to the original, unperturbed results by quantifying (i) the number of significant *L–R* pairs, (ii) the number of significant downstream branches, (iii) the count of unique downstream genes, and (iv) the change in PAS. To evaluate the performance of S2C2 relative to other cell–cell communication tools, LIANA+ and NicheNet were applied to two simulated datasets: one with simulated MAPK upregulation and another with RAS downregulation. For each tool, the number of predicted *L–R* pairs specific to the perturbed pathway was quantified.

Multicell cascading crosstalk prediction using S2C2

Signaling cascades were inferred by S2C2 to trace communication from microglia to astrocytes to excitatory neurons (Mic→Ast→Exc) and from astrocytes to microglia to excitatory neurons (Ast→Mic→Exc). Transcription factors were extracted from the Mic→Ast/Ast→Mic pathways, and their downstream targets were identified using TF–target interaction databases. Ligands that overlapped with target genes were traced to excitatory neuron pathways and annotated with PAS and *p* values. A signaling axis was revealed in which astrocytic transcription factors, including FOS and JUN, regulated ligands such as VEGFB, SPP1, and VEGFA. These ligands were found to activate downstream cascades in excitatory neurons, particularly those involved in focal adhesion signaling. Notably, FOS and JUN, key regulators in our Mic→Ast signaling cascade, were central to the ERK-FOS inflammatory axis implicated in Alzheimer's pathology.

Implementation of S2C2

To support customized implementation of S2C2, three different approaches have been developed: a GUI, CLI, and R script. The S2C2 GUI is Java-based, ensuring cross-platform compatibility across Windows, macOS, and Linux, with built-in visualization capabilities. The CLI version is designed for cloud server environments, while the R script provides a fast and efficient option tailored for bioinformaticians working with single-cell and spatial transcriptomics (ST) data in R. These options make S2C2 a versatile solution for users on various operating systems.

Parallel computing is supported by S2C2, allowing multiple calculations for selected cell type pairs to be executed efficiently. A detailed tutorial, accessible through the provided GitHub link (https://github.com/methodistsmab/S2C2/), has been created with instructions for installing S2C2 on Windows, macOS, and Linux. System requirements have been outlined, recommending at least 8 GB of RAM, with 16 GB preferable for handling large-scale single-cell and ST data, as well as a multicore processor with at least 2.5 GHz per core. For processing multiple cell–cell crosstalk pairs, four or more cores are advised to significantly enhance performance.

In the tutorial, guidance is provided on utilizing key S2C2 features, such as generating *L–R* chord diagrams and identifying significant branches. Other functionalities, including project management, progress monitoring, drug discovery tools, highlighting pathways and sub-pathways, organizing network layouts, and controlling network edge visibility and properties, are also explained. Additionally, network node detection is supported. Essential files are generated, which can be directly used with other tools, such as Cytoscape for visualization and Connectivity Map for drug repositioning.

Design and implementation of prompting strategies

Four distinct prompting strategies were implemented using the web-based version of GPT-4o: (1) zero-shot prompting, (2) zero-shot chain-of-thought (CoT) prompting, (3) few-shot prompting, and (4) few-shot CoT prompting.

- In zero-shot prompting, GPT-°was provided with simple, direct instructions that conveyed the essential information about the input file types and formats, without the inclusion of illustrative examples.
- In few-shot prompting, curated examples of validated cell–cell interactions—drawn from peer-reviewed studies across diverse disease contexts—were presented as reference points to enhance the model's ability to generalize and predict novel interactions grounded in established biological knowledge.
- In CoT prompting, the model was systematically guided through a stepwise reasoning process. These steps included (i) identification of relevant *L–R* pairs, (ii) analysis of downstream signaling pathways linked to the identified pairs, and (iii) assessment of their biological relevance in specific disease contexts. The model was explicitly instructed to "complete the task above step-by-step" to encourage structured reasoning, transparency, and interpretability.

All prompting strategies were subjected to iterative refinement to improve clarity and precision. This refinement process was

driven by the continuous evaluation of model outputs. Whenever ambiguous, incorrect, or incomplete responses were encountered, the prompts were revised to include more detailed task instructions, clearer definitions of biological concepts, and additional criteria necessary for accurate model inference. These iterative adjustments minimized ambiguity, enhanced interpretability, and improved the biological validity of the generated outputs.

Validation and interpretation of iS2C2 predictions

The prompt engineering strategies were evaluated based on three of the four robustness axes used to evaluate the iS2C2 framework, including accuracy, reproducibility, and interpretability, while excluding sensitivity to perturbation, as this metric applies only to evaluating S2C2.

- Accuracy was assessed by calculating the proportion of *L–R* pairs and their associated downstream pathways predicted by GPT-°that overlapped with those derived from S2C2-based cell–cell interaction predictions.
- Reproducibility was measured using the APRT, defined as the proportion of *L–R* pairs consistently identified in at least 80% (i.e., four out of five) of repeated runs.
- Interpretability was evaluated through a blinded assessment conducted by six domain experts specializing in cell–cell interactions and disease biology. Confidence scores were provided under three sequential conditions: (1) relying solely on domain expertise; (2) supplementing expert judgment with GPT-4o–generated hypotheses contextualized by S2C2 predictions; and (3) further incorporating independently identified literature evidence.

Interpretability ratings were recorded using a five-point Likert scale, ranging from 1 (strong disagreement with disease relevance) to 5 (strong agreement).

Evaluation of large language models for S2C2

Three LLMs (Llama 3.2, Gemini 2.5 Pro, and KIMI K2) were evaluated, and their performance was compared to previously obtained results from ChatGPT. Two types of prompt strategies were employed: simple prompts, which provided direct instructions without examples, and enhanced prompts, which included a chain-of-thought approach with illustrative examples to guide model reasoning. Llama 3.2 was assessed only in the simple zero-shot configuration.

Significant *L–R* interactions generated from single-cell RNA sequencing analyses (LLM_significant_branches.csv) were supplied to the models, which were then tasked with predicting biologically relevant *L–R* pairs and their downstream signaling pathways. Gemini 2.5 Pro and KIMI K2 were accessed through their respective web-based chat interfaces, while Llama 3.2 was operated locally via the Msty desktop application.

Model performance was assessed using two primary metrics. Accuracy was defined as the proportion of correctly predicted *L–R* pairs and downstream pathways. The APRT was calculated as the proportion of *L–R* pairs appearing in at least 80% of repeated experiments. The results from Gemini 2.5 Pro, KIMI K2, and Llama 3.2 were benchmarked against ChatGPT outcomes obtained under four conditions: simple zero-shot, simple few-shot, enhanced zero-shot, and enhanced few-shot.

Wet lab validation of S2C2 predictions

To validate the predictions of the cointelligent framework related to AD (Fig. 4b), experimental approaches were employed to assess the consistency between in silico predictions and pathological outcomes. The interactions between astrocytes and excitatory neurons were examined using in vitro models, with gene knockdown perturbations, phenotypic observations, and the assessment of proinflammatory markers. For further details, please refer to the subsequent sections in the Methods.

Astrocyte culture

Human astrocyte cell lines (ACBRI 376) were purchased from Cell Systems (ACBRI 376). Upon arrival, the cell line was stored in liquid nitrogen vapor. One day later, the cells were thawed and cultured promptly to maintain optimal viability. Culturing was performed in ATCC-formulated Dulbecco's Modified Eagle's Medium (Catalog No. 30-2002) supplemented with 10% FBS, 1% N-2 Plus Media Supplement (AR003, R&D Systems), and 1% penicillin–streptomycin (Lonza, Switzerland). After 3 days of incubation, the medium was removed, and the flasks were washed three times with D-PBS. Cells were then trypsinized with 0.05% trypsin-EDTA (Gibco, MA, USA) for 5 min at 37 °C. The enzymatic reaction was halted by adding ten volumes of DMEM, and the cells were centrifuged at $2000 \times g$ for 5 min at room temperature. The cell pellet was resuspended in 12-well plates with astrocyte media as needed. Cells were maintained at 37 °C with 5% CO_2 for 2 days before being subjected to shRNA treatment in Opti-MEM® medium. One day after transfection, the Opti-MEM® medium was replaced with fresh astrocyte media for primary astrocyte collection or Dulbecco's modified Eagle's medium (DMEM) (Gibco, MA, USA) with high glucose containing 1% penicillin–streptomycin and 10% FBS for coculture with neuronal cells.

Neural progenitor cell line culture

ReNCell VM cells (EMD Millipore, SCC008) were obtained from EMD Millipore, USA. As previously reported, ReNCell VM cells were seeded into flasks treated with Cell Basement Membrane (ATCC, ACS-3035) at a density of 4×10^3 cells/well (1.25×10^4 cells/cm^2) using ReNcell® NSC Maintenance Medium (Millipore Cat. No. SCM005) with a Growth Kit (ATCC® ACS-3003™) for neural progenitor cell expansion. The cells were maintained until reaching 90–95% confluency and then washed three times with D-PBS. Subsequently, 0.5 ml of Accutase (Millipore Cat. No. SCR005) was added, and the cells were incubated in a 37 °C and 5% CO_2 incubator for 3–5 min. After incubation, 3 ml of prewarmed ReN proliferation medium was added, and the cell suspensions were transferred to a 15 ml sterile conical tube for centrifugation at $2000 \times g$ for 3 min. After removing the supernatant, the cell pellet was resuspended in 10 ml of ReN proliferation medium. Progenitor cells were induced to differentiate by removing the growth factors bFGF and EGF from the tissue culture medium for 4-7 days. The differentiated neurons were then maintained at 37 °C and 5% CO_2 for subsequent analysis or for coculture with astrocytes and neurons.

Target gene knockdown with shRNA

Short hairpin RNAs (shRNAs) were utilized to selectively knock down genes coding for ligands and/or receptors in astrocytes and neurons, respectively. The shRNA knockdown lentiviral vectors were obtained from Vectorbuilder, with sequences provided in Supplementary Table 4. Cells were transiently transfected with the specified lentivirus using Lipofectamine LTX in combination with Plus Reagent (Thermo Fisher, Catalog number: 15338030). Target gene expression was monitored 2–3 days post-transfection. After optimizing the dosage and treatment duration of the shRNA, the shRNAs were introduced into the indicated cells to achieve gene knockdown.

Astrocyte-neuron coculture

The culture medium from the astrocytes, with or without target shRNA pretreatment, was collected to incubate the well-differentiated neurons. All ligands secreted by astrocytes were present in the medium, mimicking the *L–R* crosstalk between astrocytes and neurons in the coculture model.

Quantitative PCR (qPCR) analysis

To determine the target gene expression levels upon shRNA treatment, total RNAs were extracted using the RNeasy Plus Mini Kit (Qiagen, 74134) following the manufacturer's instructions. The quality of the RNA was evaluated using a NanoDrop 2000, and cDNA synthesis of each RNA sample was performed using the Superscript VILO cDNA Synthesis Kit (Thermo Scientific, Cat. 11754050) according to the manufacturer's instructions. Gene expression levels were quantified via qPCR using Perfecta® SYBR® Green FastMix® Reaction Mixes (QuantaBio, Reactions per Kit = 5000, Cat. 101414-280) with a LightCycler® 480 Instrument (Roche Sequencing & Life Science). The results were normalized to the *β-ACTIN* level and calculated using the $2^{-\Delta\Delta Ct}$ method. Details of the primer sequences are provided in Supplementary Table 4. Each qPCR was independently performed in triplicate.

Bone metastasis models

All animal experiments were conducted in accordance with protocols approved by the Baylor College of Medicine Institutional Animal Care and Use Committee. The bone metastasis model was established by internal iliac artery injection as previously described.[50] Briefly, mice were anesthetized and secured on a plastic board. An ~1.5 cm incision was made along the line between the femoral and iliac bones, followed by blunt dissection of the muscles to expose the common iliac artery. LLC1 cells suspended in 0.1 ml PBS were injected into the artery using a 31 G needle. Hemostasis was achieved by applying gentle pressure with a cotton tip until bleeding ceased (approximately 10 min). The incision was then sutured, and the animals were monitored until full recovery from anesthesia and restoration of normal movement. Tumor bone metastases were imaged using the IVIS system. Methods for the collection of spatial or single-cell data from mouse bone metastases have been previously reported.[95]

For in vivo treatment, tamoxifen was administered by intraperitoneal injection at a dosage of 5 mg/kg/day. For in vitro experiments, MCF-7 cells were treated with 100 nM tamoxifen for 24 h, followed by collection for qPCR analysis as described above.

Image processing

Images were acquired using a SLIDEVIEW VS200 research slide scanner (Olympus, USA). Consistent settings were uniformly applied to all images from the same experimental set. The analysis was conducted by a blinded investigator using the open-source ImageJ/Fiji software (NIH, Version 1.52a).

Comparison of existing cell–cell communication algorithms

In this study, two widely used algorithms for single-cell data analysis, CellChat[14] and LIANA+,[17] were compared with S2C2 using a public AD dataset. The outputs of S2C2, NicheNet, and LIANA+ were systematically compared based on several quantitative metrics. First, the total number of significant *L–R* pairs identified by each algorithm was quantified. Second, the functionality for downstream prediction was compared between S2C2 and NicheNet by enumerating their predicted ligand-target links. Finally, the concordance among the algorithms was determined by calculating the intersection of their respective *L–R* and LT prediction sets. This allowed for an objective assessment of the consistency and divergence in outputs across the different computational methods.

Statistical analysis

GraphPad Prism 9 and/or 10 (GraphPad Software, CA, USA) was used for all statistical analyses. Data from individual experiments are presented as the mean ± standard deviation and were assessed using Student's *t*-test, one-way ANOVA, or two-way ANOVA, followed by Tukey's post hoc test for multiple comparisons. For comparisons involving more than two groups, paired one-way ANOVA was utilized, followed by Tukey's post hoc test. Statistical significance was denoted as $*p < 0.05$, $**p < 0.01$, $***p < 0.001$, and $****p < 0.0001$.

DATA AVAILABILITY

Source code for the S2C2 crosstalk model and the iS2C2 cointelligent platform is publicly available on GitHub (https://github.com/methodistsmab/S2C2; https://github.com/methodistsmab/iS2C2). Processed single-cell and spatial transcriptomic data for the bone metastasis model, including cell-type annotations, are publicly available under accession number GSE323357. Independent validation on non-brain immune populations was performed using breast cancer and non-small cell lung cancer datasets from the Tumor Immune Cell Atlas (TICAtlas), available through Zenodo (https://doi.org/10.5281/zenodo.4263972).

ACKNOWLEDGEMENTS

This study is supported by NIH R01AG071496, NIH U01CA253553, NIH R01AG057635, NIH R01CA238727, T.T. and W.F. Chao Foundation, Cure Alzheimer's Fund, John S. Dunn Research Foundation, Houston Methodist Cornerstone Award, and the Paul Richard Jeanneret Research Fund.

AUTHOR CONTRIBUTIONS

J.T.S., L.Y., and S.T.C.W. conceptualized the research; J.T.S., L.Y., J.Y.A., Y.L.C., Z.X., S.H.Q., and M.C. conducted experiments and acquired data; J.T.S., L.Y., J.Y.A., and Y.L.C. analyzed data; J.T.S., Z.H.W., X.H.Y., and J.Y.A. developed the GUI; H.Z., Z.Y., Y.Z., Y.F.D., A.F., L.W., F.S.L., H.X.W., Z.G.J., D.X.M., S.H.Q., and W.J.D. performed the questionnaire test; M.V. contributed to imaging acquisition. A.I. improved the S2C2 computational code, and Y.Y.Z. performed cascade crosstalk prediction. D.K. provided methodological input related to Ripley's K function, and D.Y.K. evaluated AD-related results. X.H.F.Z. evaluated bone metastasis-related results. J.T.S., L.Y., J.Y.A., and S.H.Q. drafted the manuscript; J.T.S., L.Y., and S.T.C.W. reviewed and edited the manuscript. All authors have read and approved the article.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41392-026-02691-8.

Competing interests: The authors declare no competing interests.

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

REFERENCES

1. Clusmann, J. et al. The future landscape of large language models in medicine. *Commun. Med.* **3**, 141 (2023).
2. Shah, N. H., Entwistle, D. & Pfeffer, M. A. Creation and adoption of large language models in medicine. *JAMA* **330**, 866–869 (2023).
3. Mullowney, M. W. et al. Artificial intelligence for natural product drug discovery. *Nat. Rev. Drug Discov.* **22**, 895–916 (2023).
4. Liu, H., Yin, H., Luo, Z. & Wang, X. Integrating chemistry knowledge in large language models via prompt engineering. *Synth. Syst. Biotechnol.* **10**, 23–38 (2025).
5. Xianyu, Z. et al. The rise of hypothesis-driven artificial intelligence in oncology. *Cancers* **16**, 822 (2024).
6. Tian, S. et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Brief. Bioinform.* **25**, bbad493 (2023).
7. Hussein, R., Abou-Shanab, A. M. & Badr, E. A multi-omics approach for biomarker discovery in neuroblastoma: a network-based framework. *npj Syst. Biol. Appl.* **10**, 52 (2024).
8. Yang, Z., Guan, F., Bronk, L. & Zhao, L. Multi-omics approaches for biomarker discovery in predicting the response of esophageal cancer to neoadjuvant therapy: a multidimensional perspective. *Pharmacol. Ther.* **254**, 108591 (2024).
9. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
10. Pham, D. et al. Robust mapping of spatiotemporal trajectories and cell-cell interactions in healthy and diseased tissues. *Nat. Commun.* **14**, 7739 (2023).
11. Chen, H. et al. Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nat. Commun.* **10**, 1903 (2019).
12. Yeung, T. L. et al. Systematic identification of druggable epithelial-stromal crosstalk signaling networks in ovarian cancer. *J. Natl. Cancer Inst.* **111**, 272–282 (2019).

13. Choi, H. et al. Transcriptome analysis of individual stromal cell populations identifies stroma-tumor crosstalk in mouse lung cancer model. *Cell Rep.* **10**, 1187–1201 (2015).
14. Jin, S. et al. Inference and analysis of cell-cell communication using CellChat. *Nat. Commun.* **12**, 1088 (2021).
15. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* **17**, 159–162 (2020).
16. Jin, S., Plikus, M. V. & Nie, Q. CellChat for systematic analysis of cell-cell communication from single-cell transcriptomics. *Nat. Protoc.* **20**, 180–219 (2025).
17. Dimitrov, D. et al. LIANA+ provides an all-in-one framework for cell-cell communication inference. *Nat. Cell Biol.* **26**, 1613–1622 (2024).
18. Sang-Aram, C., Browaeys, R., Seurinck, R. & Saeys, Y. Unraveling cell-cell communication with NicheNet by inferring active ligands from transcriptomics data. *Nat. Protoc.* **20**, 1439–1467 (2025).
19. Almet, A. A., Tsai, Y. C., Watanabe, M. & Nie, Q. Inferring pattern-driving intercellular flows from single-cell and spatial transcriptomics. *Nat. Methods* **21**, 1806–1817 (2024).
20. Huang, L. et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* **43**, 1–55 (2025).
21. Xiong, G. et al. Toward reliable scientific hypothesis generation: evaluating truthfulness and hallucination in large language models. In *Proc. Thirty-Fourth International Joint Conference on Artificial Intelligence*, 7849–7857 (IJCAI, 2025).
22. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
23. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proc. Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 9459–9474 (Curran Associates, Inc., 2020).
24. Huang, K. et al. A foundation model for clinician-centered drug repurposing. *Nat. Med.* **30**, 3601–3613 (2024).
25. Sallam, M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare* **11**, 887 (2023).
26. Temsah, A. et al. DeepSeek in healthcare: revealing opportunities and steering challenges of a new open-source artificial intelligence frontier. *Cureus* **17**, e79221 (2025).
27. Hao, Y. et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304 (2024).
28. Knox, C. et al. DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Res.* **52**, D1265–D1275 (2024).
29. Subramanian, A. et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452.e17 (2017).
30. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).
31. Raredon, M. S. B. et al. Computation and visualization of cell-cell signaling topologies in single-cell systems data using Connectome. *Sci. Rep.* **12**, 4187 (2022).
32. Hou, R., Denisenko, E., Ong, H. T., Ramilowski, J. A. & Forrest, A. R. R. Predicting cell-to-cell communication networks using NATMI. *Nat. Commun.* **11**, 5011 (2020).
33. Baruzzo, G., Cesaro, G. & Di Camillo, B. Identify, quantify and characterize cellular communication from single-cell RNA sequencing data with scSeqComm. *Bioinformatics* **38**, 1920–1929 (2022).
34. Cabello-Aguilar, S. et al. SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Res.* **48**, e55 (2020).
35. Mathys, H. et al. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).
36. Nieto, P. et al. A single-cell tumor immune atlas for precision oncology. *Genome Res.* **31**, 1913–1926 (2021).
37. Cui, H. et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).
38. Yuksekgonul, M. et al. Optimizing generative AI by backpropagating language model feedback. *Nature* **639**, 609–616 (2025).
39. Wang, L. et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *npj Digit. Med.* **7**, 41 (2024).
40. Ferri-Borgogno, S. et al. Spatial transcriptomics depict ligand-receptor cross-talk heterogeneity at the tumor-stroma interface in long-term ovarian cancer survivors. *Cancer Res.* **83**, 1503–1516 (2023).
41. Leung, C. S. et al. Calcium-dependent FAK/CREB/TNNC1 signalling mediates the effect of stromal MFAP5 on ovarian cancer metastatic potential. *Nat. Commun.* **5**, 5092–5092 (2014).
42. Khan, O. et al. TOX transcriptionally and epigenetically programs CD8(+) T cell exhaustion. *Nature* **571**, 211–218 (2019).
43. Martinez, G. J. et al. The transcription factor NFAT promotes exhaustion of activated CD8(+) T cells. *Immunity* **42**, 265–278 (2015).
44. Goh, E. et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. *Nat. Med.* **31**, 1233–1238 (2025).
45. Hager, P. et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* **30**, 2613–2622 (2024).
46. Morabito, S. et al. Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nat. Genet.* **53**, 1143–1155 (2021).
47. Leng, K. et al. Molecular characterization of selectively vulnerable neurons in Alzheimer's disease. *Nat. Neurosci.* **24**, 276–287 (2021).
48. Olmos-Alonso, A. et al. Pharmacological targeting of CSF1R inhibits microglial proliferation and prevents the progression of Alzheimer's-like pathology. *Brain* **139**, 891–907 (2016).
49. Herdy, J. R. et al. Increased post-mitotic senescence in aged human neurons is a pathological feature of Alzheimer's disease. *Cell Stem Cell* **29**, 1637–1652.e1636 (2022).
50. Wang, H. et al. The osteogenic niche promotes early-stage bone colonization of disseminated breast cancer cells. *Cancer Cell* **27**, 193–210 (2015).
51. Zhang, W. et al. Bone metastasis initiation is coupled with bone remodeling through osteogenic differentiation of NG2+ cells. *Cancer Discov.* **13**, 474–495 (2023).
52. Quan, Y. et al. The role of N-cadherin/c-Jun/NDRG1 axis in the progression of prostate cancer. *Int. J. Biol. Sci.* **17**, 3288–3304 (2021).
53. Sun, R. et al. NTF4 plays a dual role in breast cancer in mammary tumorigenesis and metastatic progression. *Int. J. Biol. Sci.* **19**, 641–657 (2023).
54. Liu, R. et al. NTF3 correlates with prognosis and immune infiltration in hepatocellular carcinoma. *Front. Med.* **8**, 795849 (2021).
55. Bollig-Fischer, A. et al. Role of novel cancer gene SLITRK3 to activate NTRK3 in squamous cell lung cancer. *Mol. Biomed.* **2**, 26 (2021).
56. Banerjee, S. et al. CCN5/WISP-2 expression in breast adenocarcinoma is associated with less frequent progression of the disease and suppresses the invasive phenotypes of tumor cells. *Cancer Res.* **68**, 7606–7612 (2008).
57. Banerjee, S. et al. WISP-2 gene in human breast cancer: estrogen and progesterone inducible expression and regulation of tumor cell proliferation. *Neoplasia* **5**, 63–73 (2003).
58. Xu, J. et al. Interpretable deep learning translation of GWAS and multi-omics findings to identify pathobiology and drug repurposing in Alzheimer's disease. *Cell Rep.* **41**, 111717 (2022).
59. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
60. Zhang, Y., Chen, Q., Yang, Z., Lin, H. & Lu, Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci. Data* **6**, 52 (2019).
61. Birk, S. et al. Quantitative characterization of cell niches in spatially resolved omics data. *Nat. Genet.* **57**, 897–909 (2025).
62. Jonathan Thomm, G. C., Terzic, A., Hersche, M., Schölkopf, B. & Rahimi, A. In *Proc. NeurIPS 2024* (Curran Associates, Inc., 2024).
63. Nouha Dziri, X. L. et al. In *Proc. NIPS '23: 37th International Conference on Neural Information Processing Systems* (Curran Associates, Inc., 2023).
64. Binghui Peng, S. N. & Papadimitriou, C. In *Proceedings of the Conference on Language Modeling* (COLM, 2024).
65. Ge, J. et al. Development of a liver disease-specific large language model chat interface using retrieval-augmented generation. *Hepatology* **80**, 1158–1168 (2024).
66. Hao, G., Wu, J., Pan, Q. & Morello, R. Quantifying the uncertainty of LLM hallucination spreading in complex adaptive social networks. *Sci. Rep.* **14**, 16375 (2024).
67. Johnson, N. R. et al. CSF1R inhibitors induce a sex-specific resilient microglial phenotype and functional rescue in a tauopathy mouse model. *Nat. Commun.* **14**, 118 (2023).
68. Ardura-Fabregat, A. et al. Response of spatially defined microglia states with distinct chromatin accessibility in a mouse model of Alzheimer's disease. *Nat. Neurosci.* **28**, 1688–1703 (2025).
69. Baligacs, N. et al. Homeostatic microglia initially seed and activated microglia later reshape amyloid plaques in Alzheimer's Disease. *Nat. Commun.* **15**, 10634 (2024).
70. Chitu, V. et al. Microglial homeostasis requires balanced CSF-1/CSF-2 receptor signaling. *Cell Rep.* **30**, 3004–3019.e5 (2020).
71. Guan, Z. et al. Injured sensory neuron-derived CSF1 induces microglial proliferation and DAP12-dependent pain. *Nat. Neurosci.* **19**, 94–101 (2016).
72. Li, C. et al. Differential roles of astrocytic CSF1 in Alzheimer's disease and cerebral amyloid angiopathy: insights from transcriptomic analysis. *Aging Dis.* **16**, 3137–3153 (2024).
73. Hu, B. et al. Insights into the role of CSF1R in the central nervous system and neurological disorders. *Front. Aging Neurosci.* **13**, 789834 (2021).
74. Han, J. et al. Inhibition of colony stimulating factor-1 receptor (CSF-1R) as a potential therapeutic strategy for neurodegenerative diseases: opportunities and challenges. *Cell. Mol. Life Sci.* **79**, 219 (2022).

75. Gerrits, E. et al. Distinct amyloid-beta and tau-associated microglia profiles in Alzheimer's disease. *Acta Neuropathol.* **141**, 681–696 (2021).
76. Wang, C. et al. Attenuated memory impairment and neuroinflammation in Alzheimer's disease by aucubin via the inhibition of ERK-FOS axis. *Int. Immunopharmacol.* **126**, 111312 (2024).
77. Johansson, A. et al. Differential long-term tamoxifen therapy benefit by menopausal status in breast cancer patients: secondary analysis of a controlled randomized clinical trial. *J. Natl. Cancer Inst.* **117**, 868–878 (2025).
78. Bouguettaya, A., Stuart, E. M. & Aboujaoude, E. Racial bias in AI-mediated psychiatric diagnosis and treatment: a qualitative comparison of four large language models. *npj Digit. Med.* **8**, 332 (2025).
79. Bedi, S. et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA* **333**, 319–328 (2025).
80. Mesko, B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J. Med. Internet Res.* **25**, e50638 (2023).
81. Strobelt, H. et al. Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. *IEEE Trans. Vis. Comput Graph.* **29**, 1146–1156 (2023).
82. Feng, Y. et al. Knowledge graph-based thought: a knowledge graph-enhanced LLM framework for pan-cancer question answering. *Gigascience* **14**, giae082 (2025).
83. Soman, K. et al. Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics* **40**, btae560 (2024).
84. Matsumoto, N. et al. KRAGEN: a knowledge graph-enhanced RAG framework for biomedical problem solving using large language models. *Bioinformatics* **40**, btae353 (2024).
85. Xiao, Y. et al. FuseLinker: Leveraging LLM's pre-trained text embeddings and domain knowledge to enhance GNN-based link prediction on biomedical knowledge graphs. *J. Biomed. Inform.* **158**, 104730 (2024).
86. Jeyaraman, M., Balaji, S., Jeyaraman, N. & Yadav, S. Unraveling the ethical enigma: artificial intelligence in healthcare. *Cureus* **15**, e43262 (2023).
87. Gallifant, J. et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat. Med.* **31**, 60–69 (2025).
88. Preiksaitis, C. et al. The role of large language models in transforming emergency medicine: scoping review. *JMIR Med. Inform.* **12**, e53787 (2024).
89. Pesecan, C. M. & Stoicu-Tivadar, L. Increasing trust in AI using explainable artificial intelligence for histopathology—an overview. *Stud. Health Technol. Inform.* **305**, 14–17 (2023).
90. Di Martino, F. & Delmastro, F. Explainable AI for clinical and remote health applications: a survey on tabular and time series data. *Artif. Intell. Rev.* **56**, 5261–5315 (2023).
91. Allen, B. The promise of explainable AI in digital health for precision medicine: a systematic review. *J. Pers. Med.* **14**, 277 (2024).
92. Naz, Z. et al. An explainable AI-enabled framework for interpreting pulmonary diseases from chest radiographs. *Cancers* **15**, 314 (2023).
93. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
94. Handly, L. N., Pilko, A. & Wollman, R. Paracrine communication maximizes cellular response fidelity in wound signaling. *eLife* **4**, e09652 (2015).
95. Niyakan, S. et al. MUSTANG: multi-sample spatial transcriptomics data analysis with cross-sample transcriptional similarity guidance. *Patterns* **5**, 100986 (2024).