

scDock: streamlining drug discovery targeting cell-cell communication via scRNA-seq analysis and molecular docking

Chen-Hao Huang¹, Yen-Jen Oyang¹, Hsuan-Cheng Huang^{2,*}, Hsueh-Fen Juan^{1,3,4,*}

¹Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei 106, Taiwan

²Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, Taipei 112, Taiwan

³Department of Life Science, National Taiwan University, Taipei 106, Taiwan

⁴Center for Computational and Systems Biology, National Taiwan University, Taipei 106, Taiwan

*Corresponding authors. Hsuan-Cheng Huang, Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, No.155, Sec.2, Linong St., Taipei 112, Taiwan. E-mail: hsuancheng@nycu.edu.tw; Hsueh-Fen Juan, Graduate Institute of Biomedical Electronics and Bioinformatics, Department of Life Science, Center for Computational and Systems Biology, National Taiwan University, No.1, Sec.4, Roosevelt Rd., Taipei 106, Taiwan.

E-mail: yukijuan@ntu.edu.tw

Associate Editor: Anthony Mathelier

Abstract

Summary: Identifying drugs that target intercellular communication networks represents a promising therapeutic strategy, yet linking single-cell RNA sequencing (scRNA-seq) analysis to structure-based drug screening remains technically challenging and requires substantial bioinformatics expertise. We present scDock, an integrated and user-friendly pipeline that seamlessly connects scRNA-seq data processing, cell-cell communication inference, and molecular docking-based drug discovery. Through a single configuration file, users can execute the complete workflow, from raw scRNA-seq data to ranked drug candidates, without programming skills. scDock automates the identification of disease-relevant ligand-receptor interactions from scRNA-seq data and performs structure-based virtual screening against these communication targets using Protein Data Bank (PDB) or AlphaFold-predicted protein structures. The pipeline generates comprehensive outputs at each stage, enabling users to explore intercellular signaling alterations and discover therapeutic compounds targeting specific cell-cell communications. scDock addresses a critical gap by providing an accessible end-to-end solution for communication-targeted drug discovery from single-cell data.

Availability and implementation: scDock is freely available at <https://doi.org/10.6084/m9.figshare.31370368> and <https://github.com/Andrewneteye4343/scDock>. It is implemented in R, Python, shell scripts, and supports Linux systems, including Ubuntu and Debian.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) provides high-resolution gene expression profiling at the level of individual cells, offering insights into cellular heterogeneity and complex biological processes that cannot be captured by bulk RNA sequencing (Cao *et al.* 2017). Recent methodological advances, including cell-cell communication inference (Jin *et al.* 2025), transcription factor prediction (Aibar *et al.* 2017), and transcriptional dynamics modeling (Bergen *et al.* 2020), have further expanded the analytical scope of scRNA-seq, enabling a more comprehensive characterization of transcriptomic landscapes and the mechanisms that underlie diverse biological states. In particular, intercellular communication inference has emerged as a powerful approach for identifying ligand-receptor pair interactions associated with dysregulated signaling pathways in disease.

Despite offering a broader analytical repertoire than bulk RNA sequencing, scRNA-seq workflows often require complex implementation steps involving multiple programming languages and substantial computational environment setup. These technical demands pose significant challenges to many biologists and can hinder both accessibility and reproducibility. Establishing an intuitive, user-friendly analytical framework is therefore essential to lowering these barriers and improving the usability of scRNA-seq analyses.

Here, we present scDock, a command-line-based integrative toolkit that unifies core scRNA-seq analysis, intercellular communication inference, and drug screening through molecular docking of key signaling molecules. scDock accommodates multiple scRNA-seq input formats and offers accessible options for normalization, dimensionality reduction, clustering, tissue-specific annotation, and data integration. Beyond these core

Received: 21 November 2025. Revised: 20 February 2026. Accepted: 24 February 2026

© The Author(s) 2026. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

functionalities, the toolkit supports both single-group and comparative analyses of cell-cell communication inference and identifies molecules involved in differential signaling as candidate targets for molecular docking. Targeting critical components of disease-related interactomes has proven to be an effective drug discovery strategy (Lu *et al.* 2020), and scDock brings this capability into an end-to-end single-cell analysis workflow.

Despite its broad functionality, scDock remains easy to perform analytic workflow and requires only a single configuration file, which provides detailed descriptions and default settings for users with limited computational experience while allowing customization for advanced analyses. Collectively, these features position scDock as a versatile and accessible platform that streamlines single-cell transcriptomic analysis and accelerates the identification of potential therapeutic targets.

2 Implementation

scDock is implemented as an integrated toolkit that streamlines complex scRNA-seq analyses and downstream computational workflows for both human and mouse datasets. Its primary goal is to simplify the execution of scRNA-seq analysis and related bioinformatics tasks, thereby reducing the technical burden for users with limited programming experience. To achieve this, scDock is organized into three interconnected functional modules: core scRNA-seq analysis, intercellular communication inference, and molecular docking, all of which operate together to support a fully automated and customizable end-to-end workflow (Fig. 1). These modules are coordinated through R, Python, and shell scripts, enabling the entire process to be executed with a single command and a user-defined configuration file.

The configuration file specifies all key analytical parameters, including input and output paths, selected analysis methods, algorithms, and the definition of cell types or groups of interest. It provides both default and fully customizable settings, accommodating the needs of novice users while offering flexibility for advanced users. scDock can be downloaded, installed, and executed locally by following the instructions provided on Figshare and GitHub. To ensure reproducibility and cross-language compatibility, the pipeline is validated within a unified Conda environment with pinned Python and R versions, which can be readily created using the installation file provided in the Figshare and GitHub repository.

Once a scRNA-seq dataset is supplied, the pipeline automatically performs marker gene identification, cell annotation, intercellular communication inference, and molecular docking, producing outputs such as cell-annotated dimensionality reduction plot, summaries of signaling pathways, protein and compound structures, docking results, and other associated files. All intermediate data generated throughout the workflow are also saved automatically, allowing users to reuse, inspect, or extend these results for additional downstream analyses.

2.1 Core scRNA-seq analysis

In the core scRNA-seq analysis module, data are processed using the Seurat v5 framework (Hao *et al.* 2024), which performs quality control, normalization, scaling, dimensionality reduction, clustering, and integration steps when required. To ensure broad compatibility, scDock supports multiple commonly used scRNA-seq matrix formats, including 10x Genomics Cell Ranger outputs (barcodes.tsv.gz, features.tsv.gz, matrix.mtx.gz), hierarchical data format files (.h5), and plain text files (.txt). Both gene symbols and Ensembl identifiers are accepted as input formats. Since non-gene symbol identifiers can reduce annotation

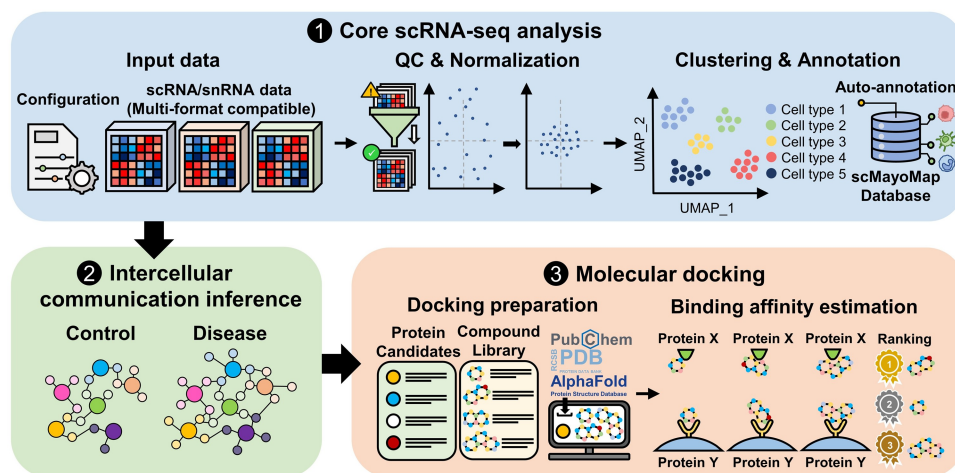


Figure 1 Overview of the scDock analytical workflow. The pipeline comprises three principal modules: the core scRNA-seq analysis module, the intercellular communication inference module, and the molecular docking module. These modules are sequentially connected, where the output generated from each module serves as the input for the subsequent module. The core scRNA-seq analysis module performs quality control, normalization, scaling, dimensionality reduction, clustering, integration, and cell type annotation. The intercellular communication inference module identifies key signaling pathways through comparison of communication probabilities across groups and selects potential therapeutic targets based on the implicated ligand-receptor pairs. In the molecular docking module, protein and compound structures are retrieved, preprocessed, and subjected to docking analysis. Candidate compounds are subsequently ranked according to their predicted binding affinities to target proteins, facilitating the identification of drugs that may modulate disease-associated ligand-receptor interactions.

accuracy and slow downstream analyses, scDock automatically converts Ensembl identifiers to their corresponding gene symbols during preprocessing. Ensembl gene IDs are converted to gene symbols using organism-specific annotation databases (org.Hs.eg.db for human, org.Mm.eg.db for mouse) via AnnotationDbi (Pagès *et al.* 2025). Unmatched IDs are retained, and duplicate symbols are suffixed numerically (e.g. SYMBOL_1) to ensure unique, Seurat-compatible feature names.

For datasets involving multiple experimental or biological conditions, users may provide a metadata file specifying sample group information. These grouping labels are automatically incorporated into a standardized field named *sample_group*, which can be used directly in downstream comparative analyses and integration workflows. During clustering and dimensionality reduction, the number of principal components (PCs) selected has a critical impact on the amount of biologically informative variation preserved in the dataset (Becht *et al.* 2018, Stuart *et al.* 2019). To determine the optimal number of PCs, scDock uses the geometric elbow method as its default strategy. This approach identifies the point with the greatest perpendicular distance from the line connecting the first and last PCs on the variance-explained plot, as previously described (Zhuang *et al.* 2022). In addition, users can manually specify the desired number of PCs in the configuration file to override the automatic selection if needed. Because batch effects can introduce unwanted technical variability that obscures true biological signals (Hie *et al.* 2019), scDock incorporates several integration methods to mitigate such artifacts, including canonical correlation analysis (CCA) (Butler *et al.* 2018), reciprocal PCA (RPCA) (Hao *et al.* 2024), and the Harmony algorithm (Korsunsky *et al.* 2019).

Manual cell type annotation is often labor-intensive, requires substantial domain expertise, and must account for tissue-specific context. To address these challenges, scDock integrates scMayoMap (Yang *et al.* 2023) for automated cell-type annotation across a broad range of tissues. Beyond the original reference database, scDock further expands the annotation resources to improve performance in disease-specific settings. These extensions currently include neuroblastoma, an uncommon and non-tissue-specific cancer that can originate in multiple organs, and breast cancer, enabling more accurate identification of malignant cells and subtype-specific populations.

2.2 Intercellular communication inference

The intercellular communication inference module uses CellChat v2 (Jin *et al.* 2025) to evaluate ligand–receptor expression patterns across cell types and infer their potential communication networks, including secreted signaling, extracellular matrix–receptor interactions, and cell–cell contact-mediated communication. Identifying ligand–receptor pairs involved in prominent signaling pathways reveals key intercellular interactions that may represent promising therapeutic targets. Communication probabilities are computed using CellChat’s permutation-based framework with 100 bootstraps, and only ligand–receptor interactions with $P < .05$ supported by ≥ 10 cells per type are retained. By default, scDock operates in a single-group mode and computes the signaling pathways with the

highest interaction probabilities across all cell types, recording both incoming and outgoing signals as well as global interaction patterns.

A central feature of scDock is its ability to investigate disease-associated signaling pathways as the initial step toward therapeutic target discovery. When a metadata file specifying experimental or biological groups is provided, scDock performs group-wise comparisons to quantify differences in intercellular communication patterns. For multi-group comparisons, ligand–receptor interactions are ranked by the absolute difference in communication probability, prioritizing condition-specific signaling while downweighting shared housekeeping interactions. Signaling pathways exhibiting the greatest shifts in interaction probabilities are identified, and their corresponding differentially active ligand–receptor pairs are designated as candidate therapeutic targets. These targets are then automatically forwarded to the molecular docking module for downstream evaluation. scDock also supports cell type-specific communication analyses. By adjusting the parameters *Run_CellChat_source_celltype* and *Run_CellChat_target_celltype* in the configuration file, users can restrict the analysis to selected source and target populations, enabling more focused and biologically informed investigations of intercellular signaling events.

2.3 Molecular docking

In the molecular docking module, scDock uses AutoDock Vina (Trott and Olson 2010) to estimate the binding affinities between small molecules and the target proteins identified through the intercellular communication analysis. Constructing compound and protein structure libraries is typically time-consuming and requires technical expertise; scDock streamlines this process by providing three options for compound library generation: (i) a built-in FDA-approved compound library, (ii) a customized library generated from Chemical Abstracts Service (CAS) registry numbers, and (iii) user-supplied compound structures.

For the first option, scDock includes a curated list of FDA-approved small-molecule drugs (version 2 September 2025) within its Figshare and GitHub repository, allowing users to rapidly assemble a ready-to-use library for systematic screening. For the second option, users may supply CAS identifiers, and scDock automatically retrieves the corresponding compound structures from PubChem (Kim *et al.* 2025) and converted them into AutoDock Vina-compatible formats using OpenBabel (O’Boyle *et al.* 2011) and RDKit (Rational Discovery LLC *et al.* 2025). For the third option, users may directly provide pre-processed compound structures that meet AutoDock Vina input specifications.

For protein structure determination, scDock provides curated reference lists (*ligand_reference.csv* and *receptor_reference.csv*) containing the best available protein models identified via UniProt-PDB-Mapper (Riziotis 2025). These structures are retrieved from the Protein Data Bank (PDB) (Berman *et al.* 2000). When no experimentally determined structures are available, scDock automatically obtains predicted models from the AlphaFold Protein Structure Database (Jumper *et al.* 2021). Users may update the curated lists when improved structures

becomes available or supply their own protein models directly. Prior to molecular docking, protein structures are preprocessed by removing non-standard residues, heteroatoms, and water molecules, adding essential hydrogens, and assigning Gasteiger charges using AutoDockTools (Morris *et al.* 2009), then converting them to PDBQT format for AutoDock Vina.

scDock performs compound screening using a global docking strategy, automatically defining the docking grid box for each protein to ensure comprehensive exploration of potential binding sites. During the docking procedure, scDock records all protein and compound structures along with the resulting predicted binding poses and binding affinities. The toolkit also summarizes and ranks the affinities of all screened compounds for each target protein, enabling efficient identification of high-priority therapeutic candidates. In addition, scDock optionally retrieves drug annotations from BindingDB (Liu *et al.* 2025) to summarize prior pharmacological evidence. A curated annotation file is available on Figshare (<https://doi.org/10.6084/m9.figshare.31370368>) and can be specified via the *Drug_Information* parameter. For each compound, known targets, organisms, affinities, and direct links to BindingDB records are retrieved.

3 Application

To demonstrate the utility of scDock, we analyzed a scRNA-seq dataset (GSE218563) from diabetic mouse kidneys (Liu *et al.* 2023), which includes samples from 16 mice representing both control and diabetic nephropathy (DN) groups. A complete walk-through of this analysis is provided in the Figshare and GitHub repository. Using the group information specified in the meta-data, we examined differences in intercellular communication between the control and DN groups. The application required approximately 50 minutes for ~27 000 cells and 20 695 genes, resulting in the identification of 14 major cell populations. Molecular docking was performed for two proteins and 43 compounds on a single CPU core with exhaustiveness set to 8, and peak memory usage reached ~14.6 GB.

Using a single configuration file, scDock produced the following results. First, it identified major kidney cell populations, including distal convoluted tubule cells, proximal tubule cells, and mesangial cells (Fig. S1, available as [supplementary data](#) at *Bioinformatics* online). In the intercellular communication analysis, scDock visualized global signaling networks and revealed pathways with increased interaction probabilities in the DN group relative to controls (File S2, available as [supplementary data](#) at *Bioinformatics* online). Among the differential signaling pathways, the interaction between peptidyl-prolyl cis-trans isomerase A (PPIA) and basigin (BSG) was highlighted as associated with DN, leading to the selection of these two proteins as targets for molecular docking.

scDock then performed molecular docking on the selected targets using a common library assembled from user-provided CAS registry numbers. The pipeline generated binding poses and calculated binding affinities for each compound-protein pair. Notably, the predicted complex between glimepiride and PPIA was visualized using PyMOL (Schrödinger LLC 2021) to provide additional structural illustration (Fig. S3, available as [supplementary data](#) at *Bioinformatics* online).

4 Conclusions

We developed scDock as an integrative and user-friendly toolkit that streamlines the complex analytical steps and computational expertise typically required for scRNA-seq data analysis. By unifying single-cell transcriptomic processing, intercellular communication inference, and molecular docking-based compound screening, scDock lowers the technical barrier for researchers and enables efficient identification of disease-associated therapeutic targets. Comprehensive documentation, installation instructions, and tutorials are available on Figshare and GitHub. To note that, molecular docking results in scDock provide computational estimates of binding affinity and structural compatibility, not direct evidence of functional effects. Docking serves to prioritize plausible drug-target interactions inferred from statistically supported cell-cell communication, which require experimental validation. We anticipate that scDock will expand the accessibility of scRNA-seq-based analyses and accelerate the discovery of potential therapeutics across a wide range of biological and disease contexts.

Acknowledgements

We thank Dr. Kai-Pu Chen at Center for Computational and Systems Biology, National Taiwan University, Dr. Ching-Chia Yang, Hong-Shian Yang, Hsin-Yo Yuan, and Hong-Ming Tseng for their valuable and constructive feedbacks.

Author contributions

Chen-Hao Huang (Data curation [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Validation [equal], Visualization [equal], Writing—original draft [equal]), Yen-Jen Oyang (Methodology [equal]), Hsuan-Cheng Huang (Conceptualization [equal], Funding acquisition [equal], Methodology [equal], Project administration [equal], Supervision [equal], Writing—original draft [equal], Writing—review & editing [equal]), and Hsueh-Fen Juan (Conceptualization [equal], Funding acquisition [equal], Methodology [equal], Project administration [equal], Supervision [equal], Writing—original draft [equal], Writing—review & editing [equal])

Supplementary material

[Supplementary material](#) is available at *Bioinformatics* online.

Conflict of interests

None declared.

Funding

This work was supported by the National Science and Technology Council [NSTC 112-2221-E-A49-061-MY3, NSTC 114-2221-E-A49-144-MY3, NSTC 113-2320-B-002-025-MY3, NSTC 113-2221-E-002-149-MY3, and NSTC 114-2321-B-002-022], Research

Proposal for NTU Core Consortiums [NTU-CC-115L892702], NTU Seed Projects for Interdisciplinary Research [NTU-SPIR-115L8417], and Center for Advanced Computing and Imaging in Biomedicine [NTU-115L900701] from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

Data availability

The scDock source code is publicly available on Figshare at <https://doi.org/10.6084/m9.figshare.31370368> and GitHub at <https://github.com/Andrewneteye4343/scDock>. The single-cell RNA-seq dataset used in the demonstration analysis was obtained from NCBI GEO database under accession number GSE218563 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE218563>).

References

- Aibar S, González-Blas CB, Moerman T *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* 2017;**14**:1083–6.
- Becht E, McInnes L, Healy J *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* 2018;**37**:38–44.
- Bergen V, Lange M, Peidli S *et al.* Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* 2020;**38**:1408–14.
- Berman HM, Westbrook J, Feng Z *et al.* The protein data bank. *Nucleic Acids Res* 2000;**28**:235–42.
- Butler A, Hoffman P, Smibert P *et al.* Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018;**36**:411–20.
- Cao J, Packer JS, Ramani V *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 2017;**357**:661–7.
- Hao Y, Stuart T, Kowalski MH *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* 2024;**42**:293–304.
- Hie B, Bryson B, Berger B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat Biotechnol* 2019;**37**:685–91.
- Jin S, Plikus MV, Nie Q. CellChat for systematic analysis of cell-cell communication from single-cell transcriptomics. *Nat Protoc* 2025;**20**:180–219.
- Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.
- Kim S, Chen J, Cheng T *et al.* PubChem 2025 update. *Nucleic Acids Res* 2025;**53**:D1516–25.
- Korsunsky I, Millard N, Fan J *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* 2019;**16**:1289–96.
- Liu S, Zhao Y, Lu S *et al.* Single-cell transcriptomics reveals a mechanosensitive injury signaling pathway in early diabetic nephropathy. *Genome Med* 2023;**15**:2.
- Liu T, Hwang L, Burley SK *et al.* BindingDB in 2024: a FAIR knowledgebase of protein-small molecule binding data. *Nucleic Acids Res* 2025;**53**:D1633–44.
- Lu H, Zhou Q, He J *et al.* Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. *Signal Transduct Target Ther* 2020;**5**:213.
- Morris GM, Huey R, Lindstrom W *et al.* AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J Comput Chem* 2009;**30**:2785–91.
- O’Boyle NM, Banck M, James CA *et al.* Open Babel: an open chemical toolbox. *J Cheminform* 2011;**3**:33.
- Pagès H, Carlson M, Falcon S *et al.* 2025. AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. R package version 1.72.0. <https://bioconductor.org/packages/AnnotationDbi>
- Rational Discovery LLC, Greg L, Julie P *et al.* RDKit (version 2025.3.6). Github. 2025. <https://github.com/rdkit/rdkit>
- Riziotis I. Uniprot-PDB-mapper (version 0.2.1). Github. 2025. <https://github.com/iriziotis/Uniprot-PDB-mapper>
- Schrödinger LLC. The PyMOL Molecular Graphics System (version 2.5.2). 2021. <https://www.pymol.org/>
- Stuart T, Butler A, Hoffman P *et al.* Comprehensive integration of Single-Cell data. *Cell* 2019;**177**:1888–902.e21.
- Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 2010;**31**:455–61.
- Yang L, Ng YE, Sun H *et al.* Single-cell Mayo Map (scMayoMap): an easy-to-use tool for cell type annotation in single-cell RNA-sequencing data analysis. *BMC Biol* 2023;**21**:223.
- Zhuang H, Wang H, Ji Z. findPC: an R package to automatically select the number of principal components in single-cell analysis. *Bioinformatics* 2022;**38**:2949–51.