NetMedPy: A Python package for Large-Scale Network Medicine Screening

Andrés Aldana,[1] Rodrigo Dorantes-Gilardi,[1] Michael Sebek,[1] Joseph Loscalzo,[2,3] Albert-László Barabási[1] and Giulia Menichetti[1,2,3,4]*

[1]Network Science Institute, Northeastern University, 360 Huntington Ave, 02115, MA, USA, [2]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, 25 Shattuck St, 02115, MA, USA, [3] Division of Cardiovascular Medicine, Brigham and Women's Hospital, Harvard Medical School, 77 Avenue Louis Pasteur, 02115, MA, USA and [4]Harvard Data Science Initiative, Harvard University, 114 Western Avenue, 02134, MA, USA

*Corresponding author. giulia.menichetti@channing.harvard.edu

Associate Editor: Name

Abstract

Summary: Network medicine leverages the quantification of information flow within sub-cellular networks to elucidate disease etiology and comorbidity, as well as to predict drug efficacy and identify potential therapeutic targets. However, current Network Medicine toolsets often lack computationally efficient data processing pipelines that support diverse scoring functions, network distance metrics, and null models. These limitations hamper their application in large-scale molecular screening, hypothesis testing, and ensemble modeling. To address these challenges, we introduce NetMedPy, a highly efficient and versatile computational package designed for comprehensive Network Medicine analyses.

Availability: NetMedPy is an open-source Python package under an MIT license. Source code, documentation, and installation instructions can be downloaded from https://github.com/menicgiulia/NetMedPy and https://pypi.org/project/NetMedPy. The package can run on any standard desktop computer or computing cluster.

Contact: Correspondence should be addressed to Giulia Menichetti, Ph.D. (giulia.menichetti@channing.harvard.edu)

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Introduction

Network medicine is a post-genomic discipline that harnesses network science principles to analyze the complex interactions within biological systems, viewing diseases as localized disruptions in networks of genes, proteins, and other molecular entities (Barabási et al., 2011). areas(do Valle et al., 2021, Nasirian and Menichetti, 2023). By integrating comprehensive biological networks, such as the interactome or protein-protein interaction network (PPI), with databases of disease-associated genes (GDA) and ligand-protein interactions, Network Medicine has: 1) successfully identified functional pathways linked to specific phenotypes and diseases (Sharma et al., 2015); 2) pinpointed potential drug targets, highlighting opportunities for drug repurposing (Cheng et al., 2018, Patten et al., 2022) and effective drug combinations (Cheng et al., 2019). Additionally, this framework has been extended beyond pharmaceuticals to identify food-derived small molecules that impact specific therapeutic areas (do Valle et al., 2021, Nasirian and Menichetti, 2023).

The structure of the biological network plays an essential role in the system's ability to efficiently propagate signals and withstand random failures.

Consequently, most analyses in Network Medicine focus on quantifying the efficiency of communication between different regions of the interactome.

For example, proteins involved in similar therapeutic areas or disease modules are expected to create a cohesive functional subgraph of proteins that communicate and influence each other. In turn, diseases with high pathobiological similarity typically reside in overlapping neighborhoods of the interactome as measured by the *separation* score (Menche et al., 2015, Supplementary Information SI).

Similarly, areas of the interactome perturbed by a drug should be close to its protein targets as quantified by the *proximity* score (Guney et al., 2016, Supplementary Information SI).

The speed and reliability of signaling are most commonly quantified through shortest-path metrics with expectations set by uniform or degree-preserving null models, highlighting biological properties not solely determined by link density or degree distribution (Supplementary Information SII). However, biological information does not always travel along geodesic paths, in part because of differences in the flow of information across links. Therefore, a comprehensive assessment of proximity and separation must consider additional metrics of diffusion and communicability. Given the complexity of these combinatorial settings, efficient algorithms are crucial for exhaustive screening of disease atlases and molecular libraries.

Despite the importance of network measures, most Network Medicine packages focus on the curation of the interactome (Helmy et al., 2022, de Carvalho, 2023) or the curation of GDAs (de Weerd et al., 2022, Ben Guebila et al., 2023). Existing packages that calculate proximity and separation have not advanced beyond their initial introduction (Wang et al., 2022, Maier et al., 2024). As a result, these tools remain highly inefficient for large-scale screening and rely exclusively on shortest-path metrics and limited sample size for hypothesis testing. Here, we introduce NetMedPy, an intuitive Python package for Network Medicine designed to quantify network localization, calculate proximity and separation between biological entities, and conduct screenings involving a large number of diseases and drugs efficiently. NetMedPy provides users with four default metrics and null models with automated statistical analyses. Optimized for high performance in large-scale studies, NetMedPy enhances the robustness and scalability of Network Medicine research, facilitating the discovery of mechanisms of action and prioritizing hypotheses for experimental validation.

NetMedPy

The workflow of NetMedPy, as illustrated in Figure 1A, involves: 1) loading the interactome, 2) computing and storing the distance matrix induced by a selected metric, 3) loading the desired GDAs and drug targets, and 4) calculating the selected scoring functions (proximity, separation) with the null models of choice. The pipeline output can be further used in downstream analyses. NetMedPy supports weighted and unweighted networks through a Graph object in NetworkX, a widely used library for network analysis. GDAs are entered using a dictionary format, where keys represent disease names and values are lists of associated genes. A similar approach is used for drug targets. The results are then returned in dictionaries, detailing the statistical analysis performed for proximity and separation. For large-scale screening studies, the output is stored in tabular form using Pandas DataFrames.

NetMedPy offers a comprehensive suite of metrics, including shortest paths (Menche et al., 2015), random walks (Masuda et al., 2017), biased random walks (Erten et al., 2011), communicability (Estrada and Hatano, 2008), and user-defined options (Supplementary Information SIII). This wide range of metrics allows researchers to tailor their analysis to the specific requirements of various biological questions. The ability to define custom metrics further empowers researchers to develop specialized approaches for their unique research needs. By

applying ensemble learning techniques, researchers can also combine the strengths of diverse metrics, enhancing the reliability and depth of their conclusions. This integrated approach can help prioritize experimental tests, improving cost-efficiency and reducing the time and effort required for validation.

NetMedPy provides primary functions for the analysis of disease modules, proximity, separation, and large-scale screening studies, including:

- Modules: Given an interactome and a set of nodes, A, NetMedPy extracts the largest connected component (LCC) or subgraph formed by set A and calculates the statistical significance of the LCC size (Supplementary Information SI.I).
- Proximity: The original proximity measure $P(A, B)$ between node sets A and B is asymmetric, meaning that $P(A, B) \neq P(B, A)$. NetMedPy addresses this property by offering both an asymmetric and symmetric proximity $P_s(A, B)$ Z-score (Supplementary Information SI.II).
- Separation: NetMedPy calculates separation (Supplementary Information SI.II) and its statistical significance, expressed by Z-Score and P-value.
- Screening: NetMedPy incorporates a screening function to calculate network measures between sets of diseases and drugs. The function runs in parallel, enhancing the computational efficiency of multi-core processing capabilities.

Network Medicine leverages null models that generate random samples as benchmarks. By comparing observed network measures against these null hypotheses, researchers can confidently assert the non-randomness of their findings, thereby substantiating the biological relevance of the observed relationships. NetMedPy enhances the robustness of this statistical analysis by incorporating various null models: Perfect Degree Match, Logarithmic Binning, Strength Binning, Uniform Distribution, and user-provided models (Supplementary Information SII). Each null model selects random node sets differently, allowing researchers to account for diverse network properties and biases that might influence the analysis.

Case Study with Vitamin D

To showcase NetMedPy, we evaluated the role of Vitamin D for an array of 13 disease phenotypes and endophenotypes, selected based on the strength of experimental evidence supporting Vitamin D as a treatment. These categories include strong support (Inflammation, Asthma, Coronary Artery Disease (CAD), Vitamin D Deficiency, Chronic Obstructive Pulmonary Disease (COPD), Rickets), medium support (Brain Neoplasms, Rett Syndrome), and low support (Prader-Willi Syndrome, Factor VII Deficiency, Beta Thalassemia, Fragile X Syndrome, Factor IX Deficiency). We curated Vitamin D's drug-target data (Piras et al., 2024) (Supplementary Information SIV.II) and the GDAs of each therapeutic area with and without experimental evidence of Vitamin D modulation (Supplementary Information SIV.III). Vitamin D is known to 1) reduce the activity of pro-inflammatory cells, 2) regulate blood pressure, and 3) reduce proliferation and boost apoptosis of cancer cells by regulating gene expression via Vitamin D receptors. Leveraging an interactome that integrates the protein-protein interactions reported in (Luck et al., 2020), (Huttlin et al. [2021]), and (Maron et al., 2021) (Supplementary
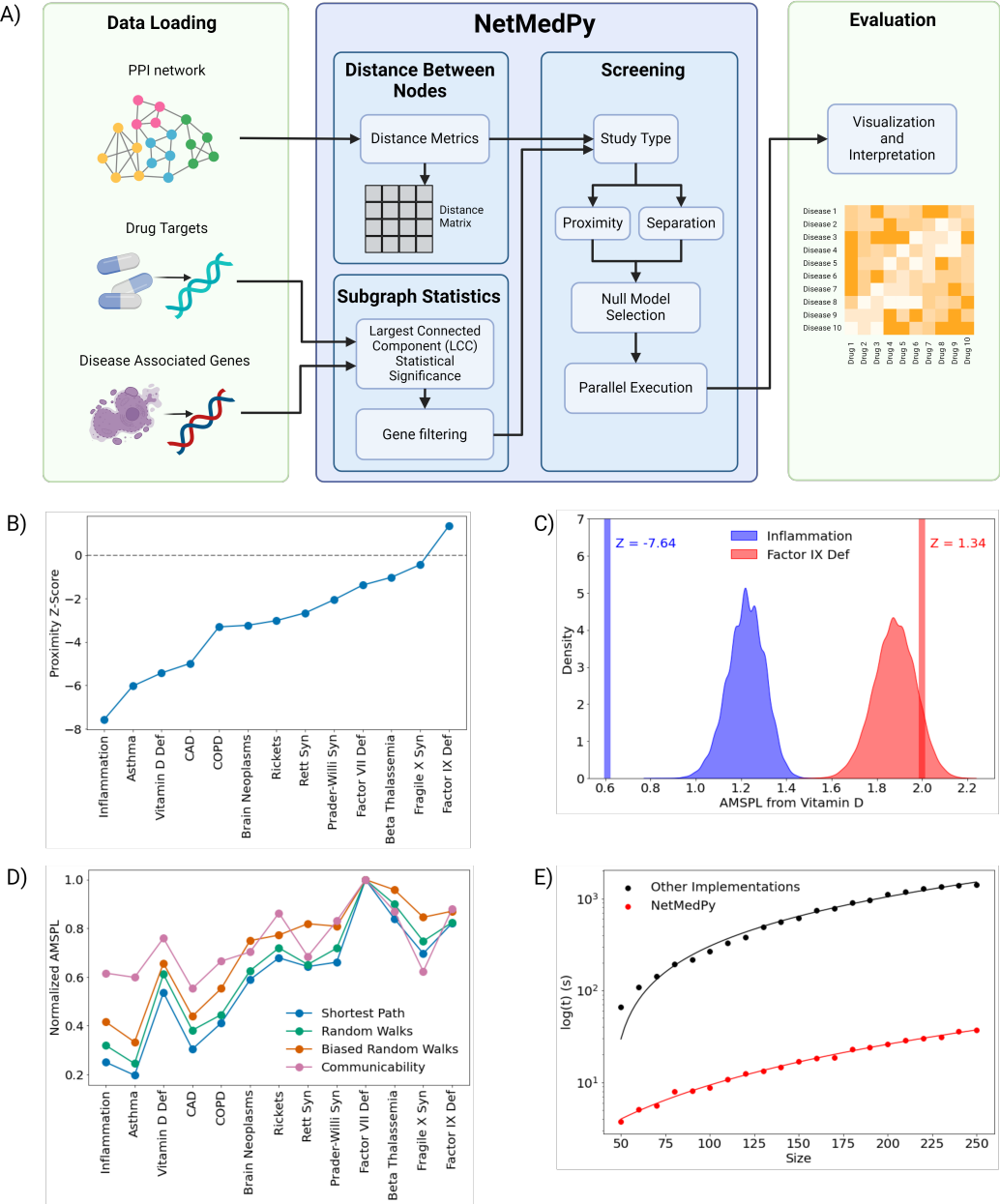
Fig. 1: **Overview and application of NetMedPy. A)** Diagram of the NetMedPy pipeline. Users first load an interaction network, drug targets, and GDAs. NetMedPy calculates the distance matrix induced by the chosen metric for all nodes in the network. Then users set options for subgraph statistics, study type (e.g., proximity, separation), null model, and execution parameters. Visualization and interpretation are performed outside of NetMedPy. **B)** Proximity between Vitamin D's targets and various diseases. A large negative Z-score indicates a statistically significant closeness between Vitamin D and the disease, while Z-scores close to zero are no different from random. **C)** AMSPL distribution and proximity Z-scores for Vitamin D to Inflammation and Factor IX Deficiency, comparing Vitamin D's targets to disease genes (vertical lines) and degree-preserving log-binned null models (density plots). Inflammation shows a significantly smaller AMSPL. **D)** Normalized AMSPL using different distance metrics: Shortest Path (blue), Random Walks (green), Biased Random Walks (orange), and Communicability (pink). **E)** NetMedPy execution time (red) versus Proximity implementations found in other packages (black; PMC11223884, PMC4740350, PMC9374494) for increasing gene set sizes. Dots represent time measurements, and straight lines indicate quadratic functions fitted to the data. In each experiment, the proximity Z-Score was calculated using one hundred random samples for illustration purposes. All calculations were performed with a 10-core Intel i9-12900H processor and 32 GB of RAM.

Information SIV.I) we calculate the proximity between Vitamin D's targets and each GDA set (Figure 1B).

Our findings reveal that the observed average minimum shortest path length (AMSPL) between Vitamin D and

inflammation is significantly smaller than expected when considering node sets of the same size and comparable degree (Z-Score = -7.64), confirming that Vitamin D influences inflammatory processes. Conversely, Factor IX Deficiency, a Mendelian disorder, is more distant from Vitamin D's targets than expected by chance (Z-Score = 1.34), providing a reasonable negative result (Figure 1B-C). When evaluating the proximity values between Vitamin D and all selected phenotypes, we find that inflammation and related diseases such as asthma show the closest proximity to Vitamin D. This result stands in contrast to diseases with no known association to Vitamin D (e.g., Prader-Willi Syndrome, Factor VII deficiency, Beta Thalassemia, Fragile X Syndrome, Factor IX Deficiency), aligning with existing literature (Figure 1B). Finally, the AMSPL-equivalents for four different metrics display a robust ranking of the results under different notions of distance (Figure 1D and Supplementary Figure S1).

NetMedPy Performance Evaluation and Comparison

Quantifying the statistical significance of network measures such as proximity and separation in large networks is computationally intensive, as it necessitates comparing selected node sets with randomly generated ones to obtain Z-scores and empirical p-values (Supplementary Information SI-SII). NetMedPy leverages parallelism and precalculated distances between all pairs of nodes to enhance performance. This optimization allows distances to be computed once and reused multiple times, significantly improving efficiency and facilitating large-scale screening studies. Figure 1E illustrates the execution time of NetMedPy for calculating proximity between random node sets of increasing size. Our findings show that NetMedPy completes this task faster than the regular proximity implementation, found in different Network Medicine packages(Wang et al., 2022, Maier et al., 2024, Patten et al., 2022), even accounting for the time required to pre-calculate the distances. Consequently, as the number of disease genes and drug-disease pairs increases, NetMedPy demonstrates a substantial performance improvement.

Discussion

We developed NetMedPy, a user-friendly Python package designed to optimize tools for Network Medicine applications. Tailored for high-performance computing, NetMedPy efficiently handles large-scale data, making it ideal for studies involving drug screening, drug repurposing, and comorbidity identification. The package offers functionalities for extracting the LCC and calculating proximity and separation between node sets, with options for both symmetric and asymmetric measures. Additionally, it supports various null models to validate the statistical significance of network metrics, ensuring robust analytical outcomes.

As in many areas of data science, the quality of predictions in NetMedPy is highly dependent on the input data. Incomplete annotations and erroneous associations can introduce variability in the results. NetMedPy enables efficient robustness analyses under perturbations to the input data (Supplementary Information SVI), and also facilitates a wide range of studies in case-specific weighted and unweighted networks. These include protein-protein interaction networks for drug repurposing (Fang et al., 2021), virus-host and drug-target networks (Zhou et al., 2020), as well as recent advances in transformers-assisted network medicine (Spector et al., 2025).

Choosing appropriate null models is essential in biological network analysis to ensure meaningful results. NetMedPy currently supports various null models, including degree-preserving node randomization. These approaches maintain the overall topology of the selected network while reassigning biological entities to different nodes. Therefore, they effectively preserve global network characteristics, such as degree-degree correlations (assortativity or disassortativity), while retaining key structural features of the selected nodes under study, including node degrees and clustering, which are essential for biologically meaningful comparisons. In biological networks, each edge carries specific biochemical significance, reflecting distinct molecular interactions and processes within the organism. In contrast, randomizing edges would disrupt this biological context, breaking meaningful associations and potentially resulting in artificially extreme p-values. Such outcomes may not accurately represent realistic biological scenarios, compromising the interpretability of the results. Nevertheless, link randomization can be valuable for assessing robustness under alternative hypotheses or for introducing explicit topological perturbations (Zhou et al., 2023). Future versions of NetMedPy may incorporate degree-preserving edge randomization to natively support such analyses.

Furthermore, the versatility of NetMedPy extends its value to numerous scientific fields that utilize networks. For example, it can enhance social network analysis by investigating social interactions and information dissemination. In epidemiology, NetMedPy can analyze disease spread and the effectiveness of health interventions within interconnected populations.

In conclusion, NetMedPy is a valuable tool for researchers, enabling them to uncover new insights and address complex problems with efficient network analysis techniques.

Supplementary data

Supplementary data are available at *Bioinformatics* online and on the GitHub repository at https://github.com/menicgiulia/NetMedPy.

Competing interests

A.-L.B. and J.L. are scientific cofounders of Scipher Medicine, Inc., which focuses on network medicine approaches to disease biomarker and drug target discovery. All other authors have no competing interests.

Author Contributions

A.A. contributed the formal analysis, methodology, validation, and visualization. A.A., M.S., G.I., and R.D-G. contributed to the methodology and the software development. A.A., M.S., G.I., and G.M. contributed to the writing and editing. J.L. and A.-L.B. contributed to the editing. G.M. conceptualized, administered, and funded the project.

Acknowledgments

We thank Ruisheng Wang for the process to curate disease-gene associations and Andrea Piras for the collection of drug-target associations.

References

A.-L. Barabási, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nat Rev Genet.*, 12:56–68, 2011.

M. Ben Guebila, T. Wang, C. M. Lopes-Ramos, et al. The Network Zoo: a multilingual package for the inference and analysis of gene regulatory networks. *Genome Biol.*, 24(1):45, 2023.

F. Cheng, R. J. Desai, D. E. Handy, et al. Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat Commun.*, 9:1–12, 2018.

F. Cheng, I. A. Kovács, and A.-L. Barabási. Network-based prediction of drug combinations. *Nat Commun.*, 10:1197, 2019.

L. M. de Carvalho. csppinet: A Python package for context-specific biological network construction and analysis based on omics data. *bioRxiv*, page 2023.05.23.541999, 2023. doi: 10.1101/2023.05.23.541999.

H. A. de Weerd, J. Åkesson, D. Guala, et al. MODalyseR—a novel software for inference of disease module hub regulators identified a putative multiple sclerosis regulator supported by independent eQTL data. *Bioinform Adv.*, 2(1):vbac006, 2022.

I. F. do Valle, H. G. Roweth, M. W. Malloy, et al. Network medicine framework shows that proximity of polyphenol targets and disease proteins predicts therapeutic effects of polyphenols. *Nat Food*, 2:143–155, 2021.

S. Erten, G. Bebek, R. M. Ewing, and M. Koyutürk. DA DA: Degree-Aware Algorithms for Network-Based Disease Gene Prioritization. *BioData Min.*, 4(1):19, 2011. doi: 10.1186/1756-0381-4-19.

E. Estrada and N. Hatano. Communicability in complex networks. *Phys Rev E.*, 77(3):036111, 2008.

J. Fang, P. Zhang, Y. Zhou, C.-W. Chiang, J. Tan, Y. Hou, S. Stauffer, L. Li, A. A. Pieper, J. Cummings, and F. Cheng. Endophenotype-based in silico network medicine discovery combined with insurance record data mining identifies sildenafil as a candidate drug for alzheimer's disease. *Nature Aging*, 1:1175–1188, 12 2021. ISSN 2662-8465. doi: 10.1038/s43587-021-00138-z. URL https://www.nature.com/articles/s43587-021-00138-z.

E. Guney, J. Menche, M. Vidal, et al. Network-based in silico drug efficacy screening. *Nat Commun.*, 7:10331, 2016.

M. Helmy, M. Mee, A. Ranjan, et al. OpenPIP: An Open-source Platform for Hosting, Visualizing and Analyzing Protein Interaction Data. *J Mol Biol.*, 434(11):167603, 2022.

E. L. Huttlin, R. J. Bruckner, J. Navarrete-Perea, J. R. Cannon, K. Baltier, F. Gebreab, M. P. Gygi, A. Thornock, G. Zarraga, S. Tam, et al. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell*, 184(11):3022–3040, 2021. doi: 10.1016/j.cell.2021.04.011.

K. Luck, D.-K. Kim, L. Lambourne, K. Spirohn, B. E. Begg, W. Bian, R. Brignall, T. Cafarelli, F. J. Campos-Laborie, B. Charloteaux, et al. A reference map of the human binary protein interactome. *Nature*, 580(7803):402–408, 2020. doi: 10.1038/s41586-020-2188-x.

A. Maier, M. Hartung, M. Abovsky, K. Adamowicz, G. D. Bader, S. Baier, D. B. Blumenthal, J. Chen, M. L. Elkjaer, C. Garcia-Hernandez, et al. Drugst. one—a plug-and-play solution for online systems medicine and network-based drug repurposing. *Nucleic Acids Research*, page gkae388, 2024. doi: 10.1093/nar/gkae388.

B. A. Maron, R.-S. Wang, S. Shevtsov, S. G. Drakos, E. Arons, O. Wever-Pinzon, G. S. Huggins, A. O. Samokhin, W. M. Oldham, Y. Aguib, et al. Individualized interactomes for network-based precision medicine in hypertrophic cardiomyopathy with implications for other clinical pathophenotypes. *Nature communications*, 12(1):873, 2021. doi: 10.1038/s41467-021-21146-y.

N. Masuda, M. A. Porter, and R. Lambiotte. Random walks and diffusion on networks. *Phys Rep.*, 716-717:1–58, 2017. doi: 10.1016/j.physrep.2017.07.007.

J. Menche, A. Sharma, M. Kitsak, et al. Uncovering disease-disease relationships through the incomplete human interactome. *Science*, 347(6224):1257601, 2015.

F. Nasirian and G. Menichetti. Molecular Interaction Networks and Cardiovascular Disease Risk: The Role of Food Bioactive Small Molecules. *Arterioscler Thromb Vasc Biol.*, 43:813–823, 2023.

J. J. Patten, P. T. Keiser, D. Morselli-Gysi, et al. Identification of potent inhibitors of SARS-CoV-2 infection by combined pharmacological evaluation and cellular network prioritization. *iScience*, 25(9), 2022.

A. Piras, S. Chenghao, M. Sebek, et al. Cpiextract: A software package to collect and harmonize small molecule and protein interactions. *bioRxiv*, page 2024.07.03.601957, 7 2024. doi: 10.1101/2024.07.03.601957.
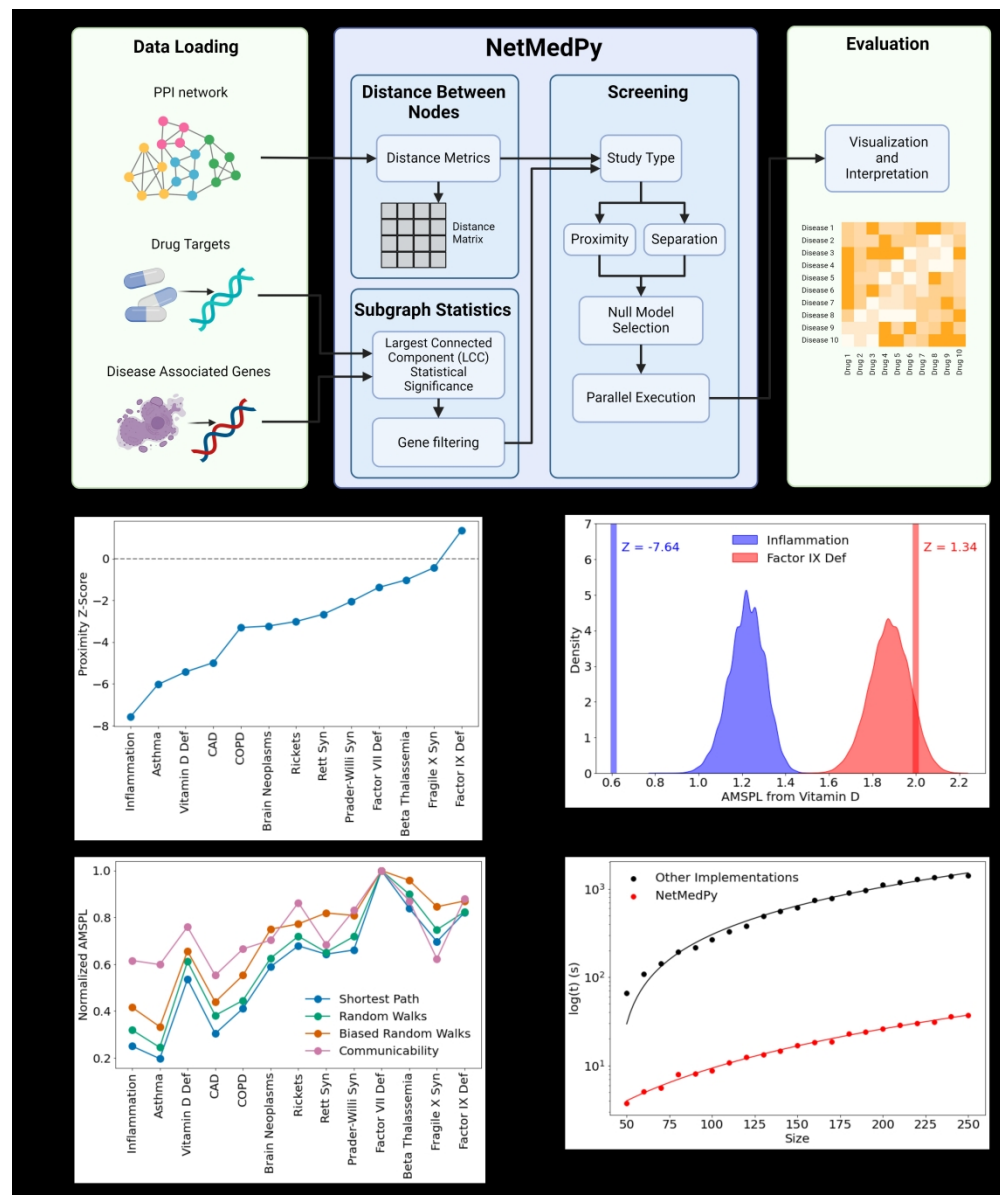
A. Sharma, J. Menche, C. C. Huang, T. Ort, X. Zhou, M. Kitsak, N. Sahni, D. Thibault, L. Voung, F. Guo, et al. A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Human molecular genetics*, 24(11):3005–3020, 2015. doi: 10.1093/hmg/ddv001.

J. Spector, A. Aldana, M. Sebek, et al. Transformers enhance the predictive power of network medicine. *medRxiv*, page 2025.01.27.25321204, 1 2025. doi: 10.1101/2025.01.27.25321204.

Y. Wang, J. Aldahdooh, Y. Hu, et al. DrugRepo: a novel approach to repurposing drugs based on chemical and genomic features. *Sci Rep.*, 12(1):21116, 2022.

Y. Zhou, Y. Hou, J. Shen, Y. Huang, W. Martin, and F. Cheng. Network-based drug repurposing for novel coronavirus 2019-ncov/sars-cov-2. *Cell Discovery*, 6:14, 3 2020. ISSN 2056-5968. doi: 10.1038/s41421-020-0153-3. URL https://www.nature.com/articles/s41421-020-0153-3.

Y. Zhou, Y. Liu, S. Gupta, M. I. Paramo, Y. Hou, C. Mao, Y. Luo, J. Judd, S. Wierbowski, M. Bertolotti, M. Nerkar, L. Jehi, N. Drayman, V. Nicolaescu, H. Gula, S. Tay, G. Randall, P. Wang, J. T. Lis, C. Feschotte, S. C. Erzurum, F. Cheng, and H. Yu. A comprehensive sars-cov-2–human protein–protein interactome reveals covid-19 pathobiology and potential host therapeutic targets. *Nature Biotechnology*, 41:128–139, 1 2023. ISSN 1087-0156. doi: 10.1038/s41587-022-01474-0.

Overview and application of NetMedPy. A) Diagram of the NetMedPy pipeline. Users first load an interaction network, drug targets, and GDAs. NetMedPy calculates the distance matrix induced by the chosen metric for all nodes in the network. Then users set options for subgraph statistics, study type (e.g., proximity, separation), null model, and execution parameters. Visualization and interpretation are performed outside of NetMedPy. B) Proximity between Vitamin D's targets and various diseases. A large negative Z-score indicates a statistically significant closeness between Vitamin D and the disease, while Z-scores close to zero are no different from random. C) AMSPL distribution and proximity Z-scores for Vitamin D to Inflammation and Factor IX Deficiency, comparing Vitamin D's targets to disease genes (vertical lines) and degree-preserving log-binned null models (density plots). Inflammation shows a significantly smaller AMSPL. D) Normalized AMSPL using different distance metrics: Shortest Path (blue), Random Walks (green), Biased Random Walks (orange), and Communicability (pink). E) NetMedPy execution time (red) versus Proximity implementations found in other packages (black; PMC11223884, PMC4740350, PMC9374494) for increasing gene set sizes. Dots represent time measurements, and straight lines indicate quadratic functions fitted to the data. In each experiment, the proximity Z-Score was calculated using one hundred random samples for

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

illustration purposes. All calculations were performed with a 10-core Intel i9-12900H processor and 32 GB of RAM.

491x584mm (236 x 236 DPI)