BIOINFORMATICS AND NETWORK MEDICINE
Sapienza University of Rome - 2024/25

Putative disease gene identification and drug repurposing for High Blood Pressure

- Ehsan Mokhtari - Mokhtari.2108539@studenti.uniroma1.it
- Arash Bakhshaee Babaroud - bakhshaeebabaroud.2105709@studenti.uniroma1.it
- Jinjia Qian - qian.2047931@studenti.uniroma1.it

Overview of the Project

Objective:

- Reconstruct the human interactome using protein-protein interaction (PPI) data.
- Integrate gene-disease association (GDA) data for disease-specific network analysis.

Tools Used:

- Python libraries: Pandas, NetworkX, Matplotlib, Seaborn.
- Data sources: BioGRID, DisGeNET, HGNC.

Outputs:

- Human interactome network.
- Disease-specific subnetwork analysis.
- Network metrics and visualizations.

Part 1.1 - Building the Human Interactome

Steps:

1. Download PPI Data:
    - Sourced from BioGRID release (BIOGRID-ALL-4.4.240.tab3.txt).
2. Filter Interactions:
    - Focused on Homo sapiens (Organism ID: 9606).
    - Kept physical interactions only.
    - Removed self-loops and duplicates.
3. Network Construction:
    - Built a graph using NetworkX.
    - Extracted the largest connected component (LCC) for further analysis.
4. Statistics:
    - Nodes in LCC: Proteins in the interactome.
    - Edges in LCC: Interactions between protein

Part 1.2 - Gene-Disease Association (GDA) Validation

Steps:

1. Load GDA Data:
    - Read curated GDAs from DisGeNET (DISEASES_Summary_GDA_CURATED_C0020538.tsv).
2. Validate Genes:
    - Compared genes with HGNC reference data (hgnc_complete_set.txt).
    - Resolved invalid genes using synonyms or alternative identifiers.
3. Save Results:
    - Validated GDA data saved to validated_gda_data.csv.
    - Validation report summarizing valid, invalid, and resolved genes saved to gda_validation_report.txt.

Outputs:

- Validated GDA file : The file is tabular data with important features such as Gene, GeneDPI, GeneDSI, GeneFullName, GenePLI, HPOClass, ...
- validation report :

```
GDA Validation Report
=====================
Total genes: 301
Valid genes: 299
Invalid genes: 2
```

Part 1.3 - Disease-Specific Interactome Analysis

Steps:

1. Disease Subgraph Creation:
   - Focused on genes associated with a specific disease (e.g., High Blood Pressure).
   - Extracted a disease-specific interactome using validated genes.
2. Largest Connected Component (LCC):
   - Isolated the LCC of the disease-specific interactome.
   - Calculated summary statistics: associated gene count, LCC size.
3. Network Metrics Computed:
   - Node Degree, Betweenness, Eigenvector, Closeness centrality.
   - Ranked top 50 disease genes based on metrics.

Outputs:

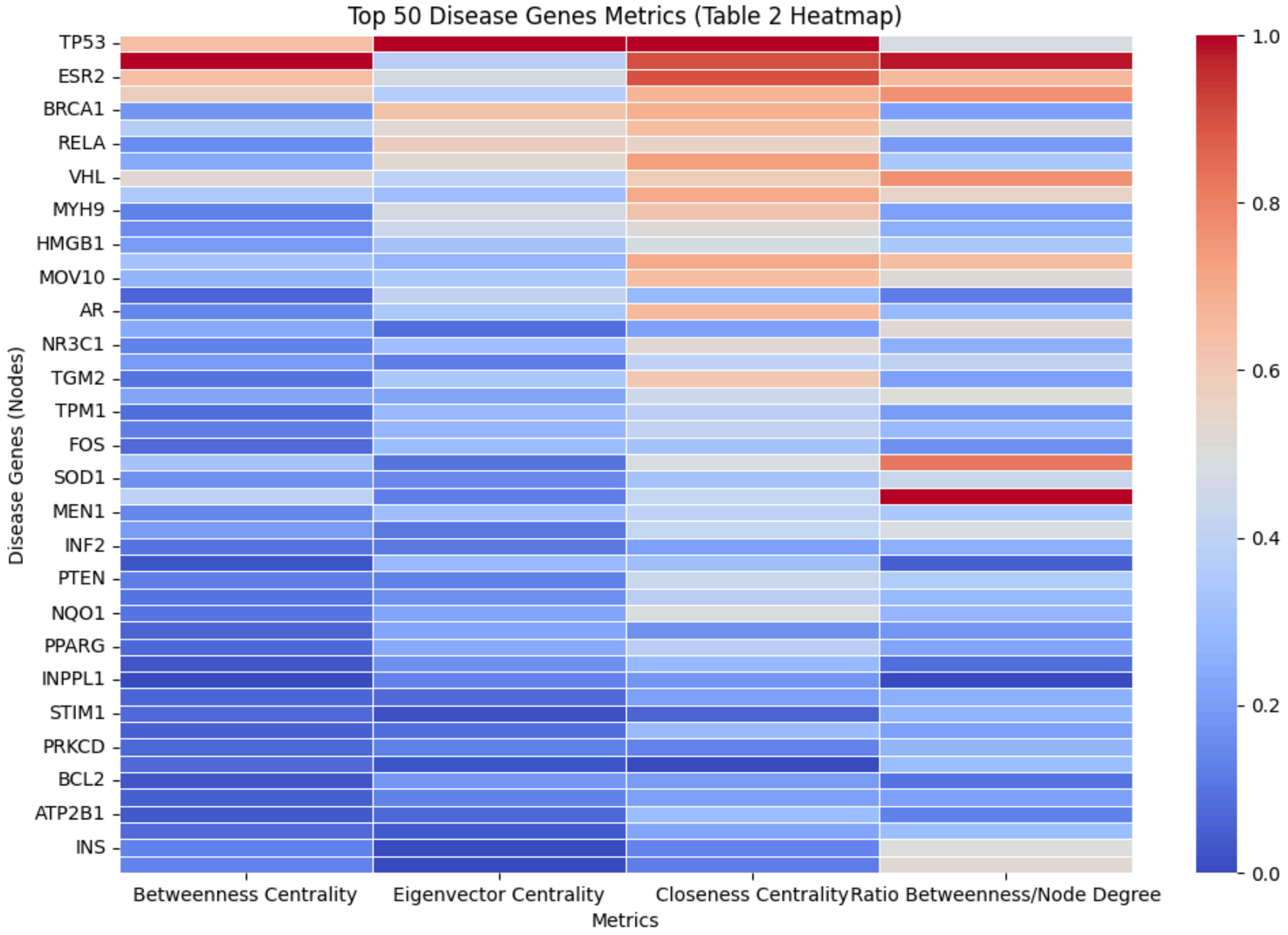
- Disease GDA summary (CSV).
- Top 50 genes ranked by metrics.

Disease Name	UMLS disease ID	MeSH disease class	Number of associated genes	number of genes present in the interactome	LCC size of the disease interactome
High Blood Pressure	C0020538	C14907489	300	286	559

Node	Node Degree	Betweenness Centrality	Eigenvector Centrality	Closeness Centrality	Ratio Betweenness/Node Degree
TP53	37	0.112787	0.356004	0.436957	0.003048
CAV1	30	0.175039	0.160273	0.425847	0.005835
ESR2	28	0.112125	0.186855	0.424947	0.004004
FN1	22	0.101800	0.153719	0.399602	0.004627
BRCA1	22	0.033648	0.236521	0.401198	0.001529
HIF1A	21	0.066906	0.203476	0.396450	0.003186
RELA	21	0.030302	0.220274	0.385797	0.001443

Visualizations and Results

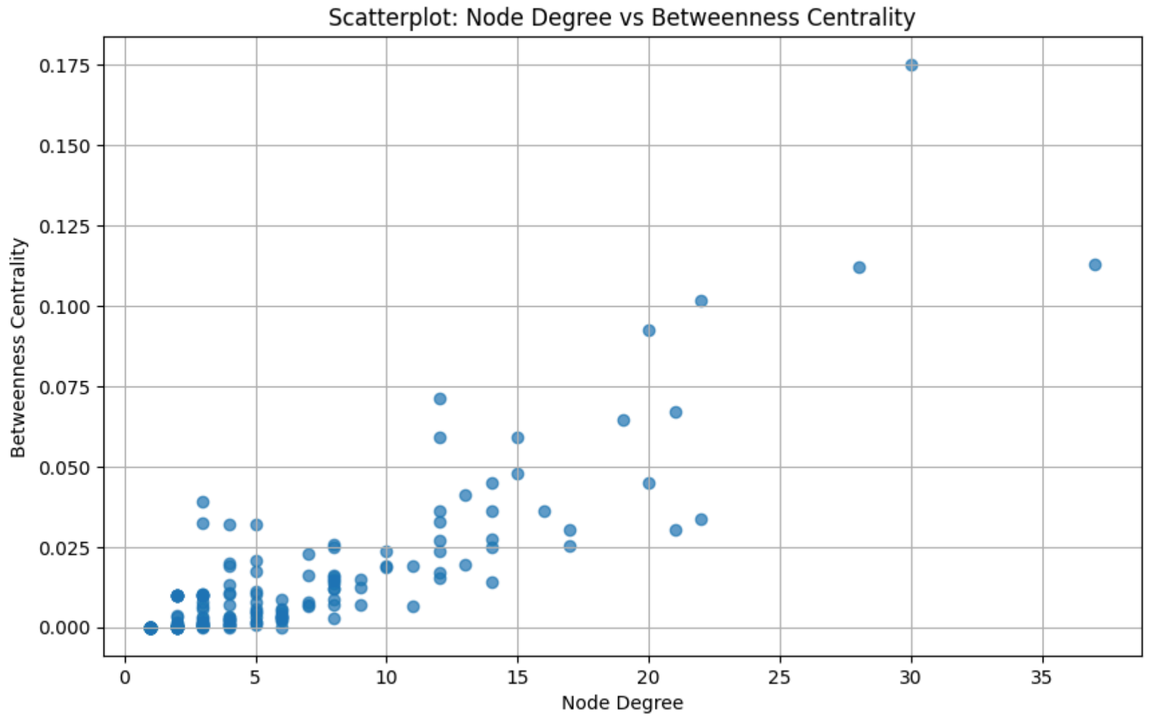
Heatmap for top 25 genes ranked by metrics from previous slide :



Visualizations and Results

Scatterplot:

- Plotted Node Degree vs. Betweenness Centrality.
- Highlighted relationships between centrality measures



Visualizations and Results

File Outputs:

- Total PPI data: totalPpi.tsv, totalppi.txt.

```
Gene1   Gene2
MAP2K4  FLNC
MYPN    ACTN2
ACVR1   FNTA
GATA2   PML
RPA2    STAT3
ARF1    GGA3
ARF3    ARFIP2
ARF3    ARFIP1
XRN1    ALDOA
APP APPBP2
APLP1   DAB1
CITED2  TFAP2A
```

Overview of Algorithmic Evaluation

- Evaluate three algorithms for disease gene identification:
  - DIAMOnD
  - DiaBLE
  - Diffusion-based model

- Use 5-fold cross-validation to assess performance

- Metrics evaluated:
  - Precision (mean ± SD)
  - Recall (mean ± SD)
  - F1-score (mean ± SD)

- Prediction performance analyzed for top 50, n/10, n/4, n/2, n genes

DIAMOnD Algorithm Evaluation

- Input:
  - PPI Network: totalppi.txt
  - Seed Genes: seed.txt
  - Parameters: top 100 predicted genes
- Output:
  - Predicted disease-related genes ranked by p-value
- Performance metrics (average ± SD):
  - Precision: 0.03 ± 0.03
  - Recall: 0.04 ± 0.03
  - F1-score: 0.03 ± 0.02

DiaBLE Algorithm Evaluation

- Input:
  - Modified DIAMOnD parameters (adjusted hypergeometric universe size).
  - PPI Network: totalppi.txt
  - Seed Genes: seed.txt
- Output:
  - Predicted disease-related genes stored in Disease_DIABLE.txt
- Performance metrics (average ± SD):
  - Precision: 0.03 ± 0.03
  - Recall: 0.03 ± 0.02
  - F1-score: 0.03 ± 0.03

Diffusion-Based Model Evaluation

- Input:
  - PPI Network: totalppi.txt
  - Seed Genes: seed.txt
  - Personalized PageRank as the diffusion model
  - Parameters: alpha = 0.85 (diffusion coefficient)
- Cross-Validation:
  - Remove seed genes from predicted rankings
- Performance metrics (average ± SD):
  - Precision: 0.01 ± 0.02
  - Recall: 0.03 ± 0.03
  - F1-score: 0.01 ± 0.01

Comparative Summary of Algorithms

Algorithm	Precision	Recall	F1-score
DIAMOnD	0.03 ± 0.03	0.04 ± 0.03	0.03 ± 0.02
DiaBLE	0.03 ± 0.03	0.03 ± 0.02	0.03 ± 0.03
Diffusion	0.01 ± 0.02	0.03 ± 0.03	0.01 ± 0.01

- DIAMOnD and DiaBLE showed comparable performance.
- Diffusion-based model underperformed in all metrics compared to DIAMOnD and DiaBLE

Which Algorithm is Better ?

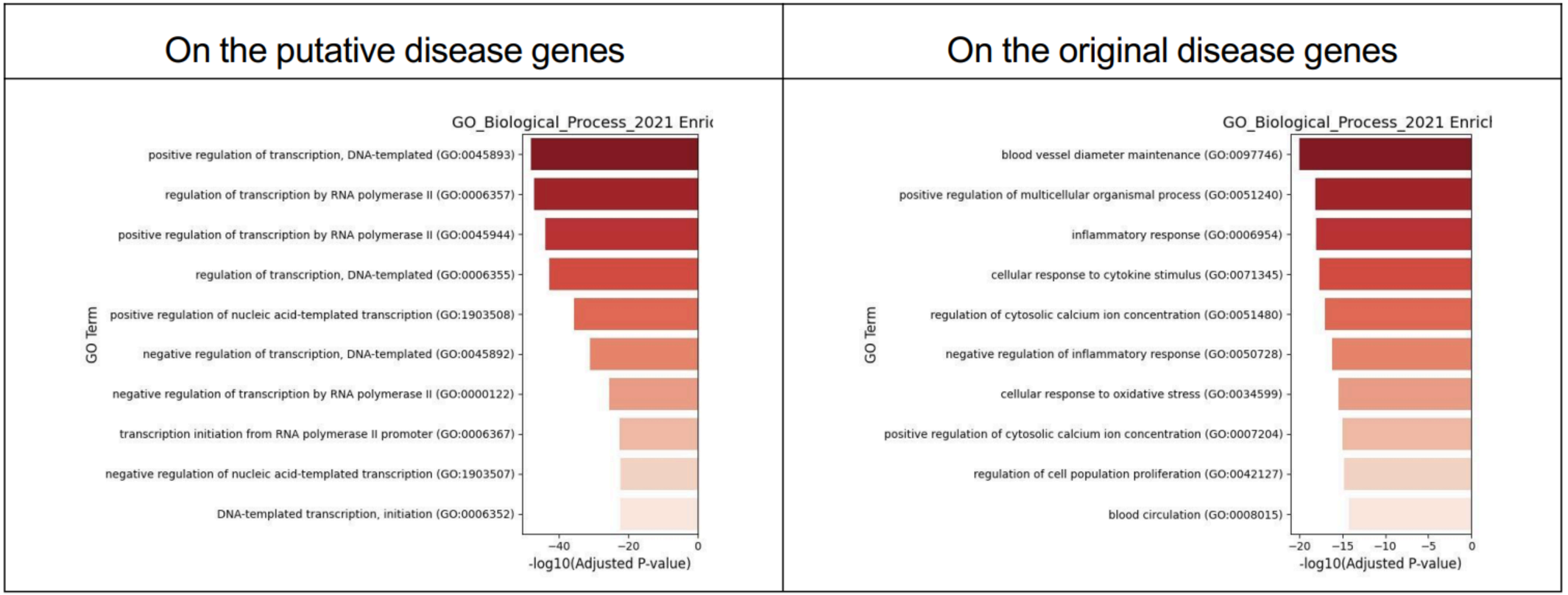
There is no universally superior algorithm!

- DIAMOnD excels at identifying putative disease genes in densely connected network regions by iteratively expanding from seed genes based on interactome structure.
- DIABLE performs better in scenarios with sparse gene-disease associations or when integrating multidimensional biological data, making it more suitable for complex datasets.

Both DIAMOnD and DIABLE exhibit similar performance metrics, with only minor differences in precision, recall, and F1-score averages and standard deviations, indicating no significant overall advantage for either.
However, DIAMOnD was slightly better and faster in processing, leading to its preference in our study.

Perform the EA over the putative and original disease genes :

We analyzed the putative and original disease genes using EnrichR to identify pathways or terms that are statistically overrepresented in five specific categories: GO-BP (biological processes), GO-MF (molecular functions), GO-CC (cellular components), Reactome pathways (biological pathways), and KEGG pathways (gene networks and metabolic pathways). Here is an example of identifying overrepresented biological processes among putative and original disease genes :

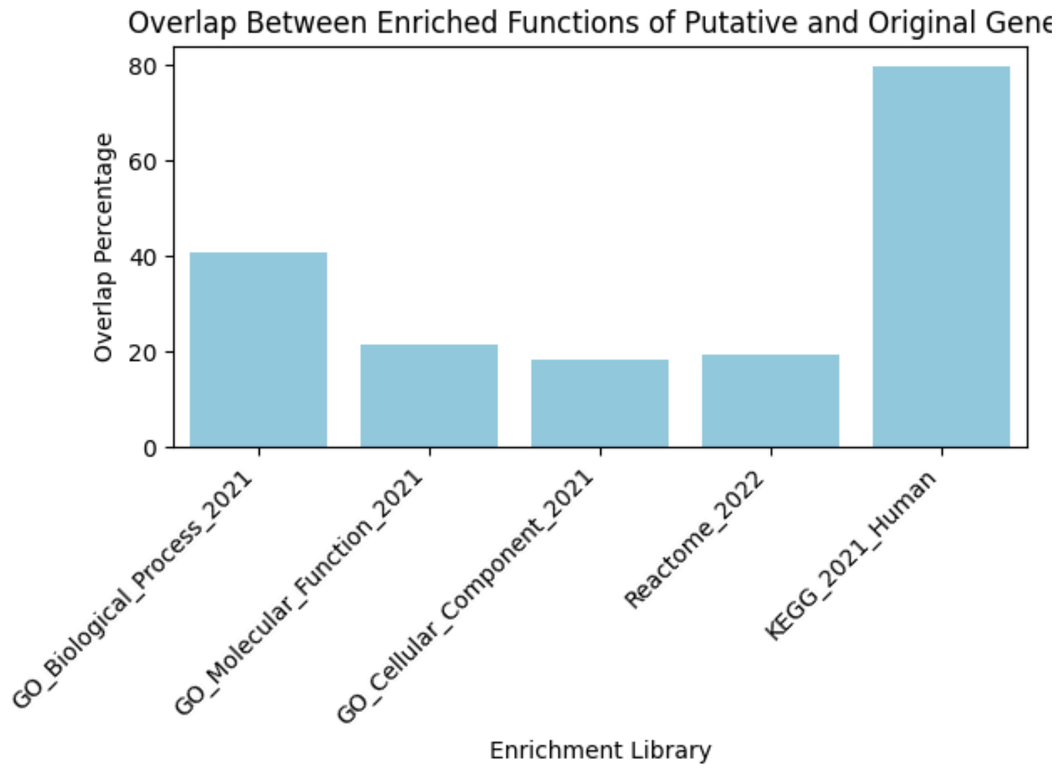


The −log10 transformation is used to improve readability, spread small p-values across a wider range for better comparison.

Overlap between enriched functions of original and putative disease gene :

comparing the enriched functions identified for original disease genes and putative disease genes, focusing on those with an adjusted p-value less than 0.05, to determine if they share any common functional or biological pathways.

The overlap between enriched functions of original disease genes and putative disease genes is as below :



Drug repurposing

We compiled a ranking of identified drugs, starting with the drug associated with the most of the above 20 genes and below is the top 3:

rank	drug_name	Gene_count
1	BORTEZOMIB	8
2	CAPIVASERTIB	7
3	CARFILZOMIB	7

Upon reviewing the top 2 drugs identified in Part 4.1 through the ClinicalTrials.gov database, we could not find any clinical trials testing the drugs for high blood pressure.

For the third one, CARFILZOMIB, we find 2 trails:

- Cardiovascular Complications of Carfilzomib (NCT04407858)
- Right Ventricular Function in Patients Taking Carfilzomib (NCT06568952)

Possible Reasons for not finding any clinical trials testing for previous drugs

- False Positives in Network Analysis : The algorithms (DIAMOnD, DIABLE) used for predicting disease-associated genes and drugs might have mistakenly linked these drugs to hypertension.
- Bias in Input Data : If the interactome or gene-disease association data were incomplete, biased, or noisy, it could affect the accuracy of the predictions.
- Drug-Disease Relationship Complexity : Just because a drug targets genes associated with a disease doesn’t guarantee it will be clinically effective for treating it.
- Lack of Research Focus : Researchers and pharmaceutical companies may not have explored these drugs for high blood pressure treatment.
- Clinical Trial Limitations : Even if some trials exist, they might not use the exact terms searched or may be unpublished or in early phases.
- Market Factors : Drug repurposing efforts depend on funding, approval processes, and commercial interest, which might not favor hypertension.


Thank
You