

# Leakage-aware retrospective validation of a GWAS-informed drug repurposing pipeline for IgA nephropathy

Hanif Rahman

hanif@hanifrahman.com

Indepent Researcher <https://orcid.org/0009-0004-4270-4222>

---

## Research Article

**Keywords:** IgA nephropathy, drug repurposing, data leakage, retrospective validation, network proximity, graph neural network, knowledge graph

**Posted Date:** October 6th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-11243586/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** The authors declare potential competing interests as follows: The author has filed two UK patent applications relevant to this work: GB2613032.8 (filed 4 June 2026), covering second medical use claims for candidates including nintedanib, dasatinib, bosutinib, and F351 in IgA nephropathy, arising from the pipeline described here; and GB2613961.8 (filed 16 June 2026), covering a target-prioritisation algorithm applied in a companion study and not the subject of this manuscript. Neither filing restricts publication of the methods or results.

---

# Leakage-aware retrospective validation of a GWAS-informed drug repurposing pipeline for IgA nephropathy

Hanif Rahman\*

Independent Researcher, London, United Kingdom

\*Corresponding author: [hanif@hanifrahman.com](mailto:hanif@hanifrahman.com)

ORCID: <https://orcid.org/0009-0004-4270-4222>

## Abstract

Retrospective recall of approved drugs is the standard validation for computational drug repurposing, but it is exposed to a leakage mode that is rarely checked: drug-target annotations in public reference databases can be added after a drug's approval, so a pipeline appears to recover a target it could not have known at the time. We describe a leakage-aware validation protocol that credits a recovered drug only when the target annotation it is scored through is present in a dated 2020 release of the reference database, and that evaluates graph neural network recall only on held-out drugs. We apply it to IgA nephropathy, the leading cause of kidney failure in East Asia, where five drugs have been approved since 2021 and no computational repurposing study has been published. A four-method pipeline combines network proximity on a 953-protein module built from 30 genome-wide association loci, a relational graph convolutional network on PrimeKG, Open Targets genetic association scoring, and LINCS L1000 transcriptomic reversal. The protocol removed the pipeline's top-ranked apparent success, sparsentan, as post-approval annotation backfill; both annotation-safe known drugs, atrasentan and fostamatinib, ranked in the top 10% of 1,989 scored compounds. The consensus nominates fifteen candidates with no trial record in this disease. Four methods converge on LYN kinase inhibitors, led by nintedanib, and a transforming growth factor beta inhibitor, F351, addresses the fibrosis axis that no approved therapy reverses.

## Introduction

IgA nephropathy is the most common primary glomerulonephritis in the world and the most frequent cause of kidney failure in East and Southeast Asia. In Japan, incidence reaches 45 per million per year. Patients of East Asian ancestry carry a hazard ratio of 2.27 for progression to end-stage kidney disease, compared with European patients<sup>1</sup>. From December 2021 to November 2025, the FDA approved five mechanism-targeted agents: budesonide, sparsentan, iptacopan, atrasentan, and sibeprenlimab<sup>2</sup>. Yet no approved therapy reverses established fibrosis or regenerates lost nephrons, and approximately 20-30% of patients in unselected cohorts progress to end-stage kidney disease within 20 years of diagnosis<sup>1</sup>.

The computational biology of IgA nephropathy has not kept pace. A 2024 meta-analysis of machine learning in the disease identified 47 published studies: 12 addressed diagnosis, 27 prognosis, and 8 other tasks. None addressed drug discovery<sup>3</sup>. Network proximity methods and graph neural networks have been applied to drug repurposing across a range of diseases, but no published study has applied them to any glomerular disease<sup>4</sup>.

This gap is not due to a shortage of data. A 2023 genome-wide association study identified 30 genome-wide significant susceptibility loci for this disease across 10,146 cases and 28,751 controls, including 16 novel associations<sup>1</sup>. The PrimeKG biomedical knowledge graph encodes 129,375 nodes and 4,050,249 edges from 20 biomedical databases<sup>5</sup>. The Drug Repurposing Hub contains 7,540 annotated compounds<sup>6</sup>. Open Targets links genome-wide-association loci to approved drugs<sup>7</sup>.

A disease with several recent approvals is also a natural test bed for retrospective validation, and here a methodological problem appears. Repurposing pipelines are usually validated by asking whether they rank already-approved drugs highly. That test assumes the drug-target annotations used for scoring were available before the approval. Public reference files are revised, and a drug that becomes prominent in a disease can gain a target annotation afterwards. A pipeline scored on a current file can then recover the drug through an annotation that did not exist at the time it claims to be predicting from. Graph neural network recall has a parallel failure: a known drug that sits in the training split is fitted, not predicted. Neither failure is visible in the headline recall number.

We present a four-stage pipeline (Figure 1): disease module construction from genome-wide-association loci, network proximity screening, genetic association cross-join, and graph neural network link prediction, with transcriptomic reversal and structural co-folding as downstream checks. Its validation applies an explicit leakage-aware protocol: known drugs count toward recall only when the target annotation they are scored through is present in a dated 2020 release of the reference database, and neural recall is measured by leave-one-out. We report one instance where the protocol overturned an apparent success, then nominate candidates for experimental follow-up. Of the 30 susceptibility loci, at least four encode proteins that are targets of approved drugs in other diseases but have no trial record in IgA nephropathy: LYN, TGFB1, REL, and FCAR.

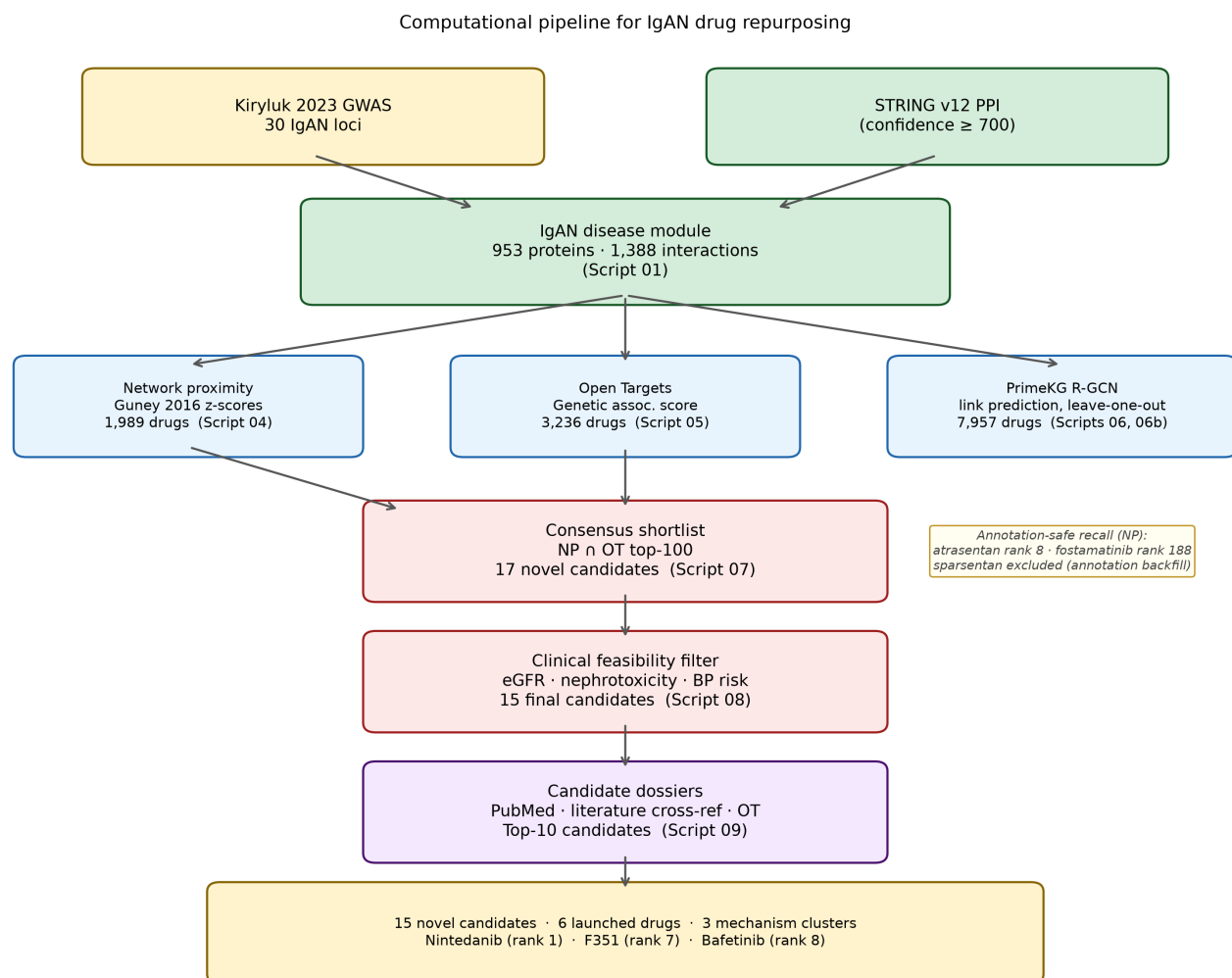


**Figure 1.** Computational pipeline. Inputs flow from 30 genome-wide-association loci and a protein interaction network into the disease module, which feeds three independent scoring methods. The consensus shortlist and clinical feasibility filter produce 15 final candidates.

## Methods

### Data sources

Genome-wide-association loci. Thirty genome-wide significant susceptibility loci from a 2023 genome-wide association study were manually curated from its supplementary tables<sup>1</sup>. Each locus was assigned a four-hit pathogenesis alignment (Gd-IgA1 glycosylation, autoantibody formation via the BAFF/APRIL axis, complement activation, fibrosis and mesangial damage) following the established four-hit hypothesis of disease pathogenesis.

Protein-protein interaction network. STRING v12 interactions for Homo sapiens (taxon 9606) were queried via the STRING REST API<sup>8</sup>. Only interactions with combined confidence score at least 700 were retained.

Drug Repurposing Hub. Drug-target annotations for scoring were taken from the Broad Institute’s drug information file with file date 18 August 2025 (7,540 compounds with mechanism-of-action and target annotations)<sup>6</sup>. The release dated 24 March 2020 was also retrieved and used only to check annotation vintage (see Leakage-aware validation protocol). Novel candidates are identified by network proximity and genetic association rather than by indication label.

Genetic association data. Gene-disease associations for the disease (EFO\_0004194) were retrieved via the Open Targets Platform GraphQL API v4<sup>7</sup>. Drug-target annotations and known clinical indications were queried for each candidate drug.

Knowledge graph. The heterogeneous biomedical knowledge graph PrimeKG was downloaded from Harvard Dataverse<sup>5</sup>. Node, edge, and drug-feature files were used with SHA256 checksums recorded.

### **Disease module construction**

The 30 loci gene symbols served as seed proteins. For each seed, all interaction partners above the confidence threshold were retrieved from the STRING v12 interaction-partners endpoint with no limit on partner count; seed-seed interactions were retrieved separately. All edges were deduplicated and an undirected graph was constructed with seed proteins annotated by hit alignment.

### **Network proximity analysis**

The z-scored network proximity measure of Guney et al.<sup>4</sup> was implemented:  $d(S,T)$  is the mean over drug target set  $S$  of the minimum shortest-path distance to the genome-wide-association seed set  $T$  in the disease module graph, and the z-score is computed against a null distribution of 1,000 degree-matched random node sets of size  $|S|$ . Of 5,499 Drug Repurposing Hub compounds with target annotations, 1,989 had at least one target in the module and were scored.

### **Leakage-aware validation protocol**

Two rules govern every retrospective recall figure reported here. First, annotation vintage: a known drug is credited as recovered by network proximity only if every target through which it is scored (its annotated targets present in the disease module) is already annotated in the 24 March 2020 Drug Repurposing Hub release. Drugs that fail this test are scored and reported, but excluded from the recall claim. Atrasentan is scored through EDNRA and fostamatinib through SYK, both annotated in the March 2020 release (fostamatinib’s later FLT3 annotation lies outside the module and does not enter its score). Sparsentan’s target field is empty in the March 2020 release; its AGTR1 and EDNRA annotation appears only in the 2025 file. Second, held-out evaluation: graph neural network recall is reported only under leave-one-out cross-validation, so each known drug is ranked by a model that never saw it as a positive example. An earlier single random split placed two of the approved drugs in the training set, and that design was discarded for this reason.

### **Graph neural network link prediction**

PrimeKG edges of relevant types were filtered to a subgraph of 35,629 nodes and 693,524 edges. Node features combined a three-dimensional type one-hot vector with z-normalised molecular descriptors for drug nodes and a 32-dimensional learnable embedding. A two-layer relational graph convolutional network encoder<sup>11</sup>, assigning one weight matrix per relation type, produced node embeddings; a two-layer multilayer-perceptron decoder predicted link probability between each drug node and the disease node. Fifteen known drugs for this disease served as positive training examples, with 7,942 other drug nodes as the negative pool (three-times oversampling per batch), trained for 200 epochs (Adam, learning rate 0.001, hidden dimension 128, dropout 0.3, binary cross-entropy loss). Validation used leave-one-out cross-validation: the model was trained 15 times, each time excluding exactly one known drug from the positive set, and the held-out drug’s rank among all 7,957 drug nodes was recorded, avoiding the circular evaluation produced by a single random split. An

initial relation-agnostic GraphSAGE encoder<sup>10</sup> was implemented first and is reported as a head-to-head architecture ablation against the relation-aware encoder on identical data.

## Consensus shortlist and clinical feasibility filter

The top-100 drugs from network proximity were intersected with the top-100 drugs from genetic association score. Two hard-removal criteria were then applied: non-therapeutic compounds, and zero genetic association score with no network proximity evidence. A clinical priority score combined phase, nephrotoxicity risk, and hypertension risk, with a bonus for candidates addressing two pathological hallmarks simultaneously.

## Transcriptomic reversal analysis

A disease signature (200 up-regulated and 200 down-regulated genes by Mann-Whitney U test) was derived from a public glomerular biopsy dataset<sup>15</sup> (20 cases, 22 controls) and submitted to Enrichr<sup>16</sup> LINCS L1000 chemical perturbation libraries in both directions; a connectivity score was computed as the geometric mean of the best combined score across both directional queries.

## Protein-ligand structure prediction

Protein-ligand co-folding was performed for the five top-ranked consensus candidates using an open-weight structure prediction model<sup>17</sup>, with three independent diffusion samples per pair and protein sequences from UniProt. The ligand interface predicted TM-score quantifies confidence in the predicted binding interface. Contact residue analysis considered protein atoms within 4 angstroms of ligand heavy atoms.

# Results

## The genome-wide-association coverage map

Of the 30 susceptibility loci (Figure 2), 11 encode proteins targeted by FDA-approved or Phase 3 drugs: TNFSF13 (sibeprenlimab), TNFSF13B and TNFRSF13B (atacept, telitacept), CFB (iptacopan), EDNRA (atrasentan), AGTR1 (sparsentan), C5 (ravulizumab), MASP2 (narsoplimab), and the MHC class II locus HLA-DQB1 (addressed indirectly by budesonide). This leaves at least four high-confidence loci with no drug in any clinical stage for this disease: LYN (rank 22 of 3,005 gene associations), REL (rank 23), TGFB1 (rank 34), and FCAR (rank 53).

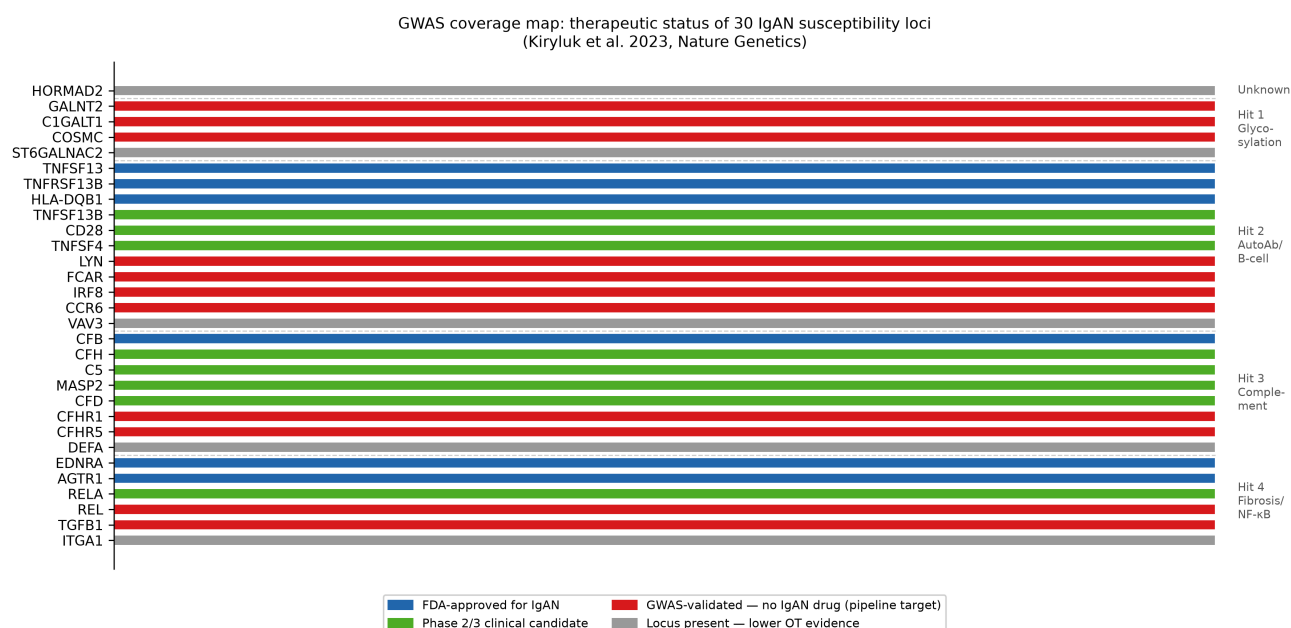


**Figure 2.** Genome-wide-association coverage map showing therapeutic status of all 30 susceptibility loci for this disease. Blue indicates an FDA-approved drug exists; green indicates a Phase 2/3 clinical candidate; red indicates a genetically validated locus with no drug for this disease; grey indicates a locus with lower genetic evidence.

The glycosylation axis (GALNT2, C1GALT1, COSMC) ranks much lower in genetic association evidence (GALNT2 at rank 1,013) because its disease-gene association derives primarily from mechanistic studies rather than large-scale functional genomics. This locus is flagged for dedicated screening using single-cell-informed network analysis in future work.

A second structural finding: the approved drugs sparsentan (targets EDNRA, AGTR1) and atrasentan (target EDNRA) have zero genetic association score, meaning their targets were not identified by the genome-wide association study. They were discovered through mechanism-driven clinical analogy from cardiorenal physiology rather than genetics. This confirms that the genetic-proximity pipeline and empirical clinical research are complementary rather than overlapping.

### Disease module construction

The 30 loci were mapped onto a protein-protein interaction network (STRING v12<sup>8</sup>, confidence threshold 700), producing a 953-protein, 1,388-interaction module. All four pathological axes of the disease are represented: the B-cell receptor and autoantibody axis (LYN, degree 136; TNFSF13B, degree 43), the fibrosis and NF- $\kappa$ B axis (RELA, degree 198; TGFB1, degree 190), the complement axis (CFH, degree 45; CFB, degree 24), and the glycosylation axis (C1GALT1, degree 33). The hub degree of RELA and TGFB1 reflects their centrality in the human interactome, and drugs whose targets connect directly to these hubs receive strongly negative proximity z-scores.

The B-cell receptor cascade is intact in the module: LYN connects to SYK and BTK as first-degree neighbours. Fostamatinib, a SYK inhibitor in Phase 2 trials for this disease, has its target one hop from LYN, predicting its network proximity rank before any scoring was performed.

### Retrospective validation by network proximity

Network proximity scored 1,989 Drug Repurposing Hub compounds against the 30 genome-wide-association seeds, using 1,000 degree-matched permutations per drug (Table 1). Two known drugs for this disease are scored through targets already annotated in the March 2020 release: atrasentan (EDNRA) and fostamatinib (SYK, one hop downstream of LYN). Both rank in the top 10% of scored compounds: atrasentan 8th ( $z = -6.00$ ) and fostamatinib 188th (9.5th percentile;  $z = +0.16$ , so its placement is rank-based rather than a significant proximity signal). A third approved drug, sparsentan (EDNRA and AGTR1), also ranks highly in the current dataset (1st,  $z = -8.37$ ), but this is a hindsight artefact rather than a retrospective success: its target field is empty in the March 2020 Drug Repurposing Hub release, and its AGTR1 and EDNRA annotation appears only in the 2025 file used for scoring. It is reported for completeness but excluded from the validation claim, since a pipeline cannot be credited with identifying a target it did not yet know. Budesonide and hydroxychloroquine, whose mechanisms are not encoded by protein-protein proximity to the genetic loci, scored at the 35th and 51st percentiles respectively, in line with expectation.

An independent check emerged without any disease-specific design: triptolide, used informally for this disease in China<sup>9</sup>, ranked 18th ( $z = -5.59$ ), recovered from NF- $\kappa$ B network topology alone, without any knowledge of its clinical use.

**Table 1.** Network proximity rankings for the top 25 novel Drug Repurposing Hub compounds scored against the disease module, excluding known drugs for this disease.

| Rank | Drug       | z-score | Dist. | Primary target | Hit | Phase       | Trial?            |
|------|------------|---------|-------|----------------|-----|-------------|-------------------|
| 2    | fr-139317  | -7.84   | 0.00  | EDNRA          | 4   | Preclinical | No                |
| 3    | saralasin  | -7.39   | 0.00  | AGTR1          | 4   | Preclinical | No                |
| 4    | fimasartan | -7.00   | 0.00  | AGTR1          | 4   | Launched    | No (ARB analogue) |

| Rank | Drug                         | z-score | Dist. | Primary target | Hit | Phase       | Trial?               |
|------|------------------------------|---------|-------|----------------|-----|-------------|----------------------|
| 5    | adaprev                      | -6.52   | 0.00  | TGFB1          | 4   | Phase 3     | No                   |
| 6    | D-4476                       | -6.38   | 0.00  | TGFB1          | 4   | Preclinical | No                   |
| 9    | candesartan                  | -6.00   | 0.00  | AGTR1          | 4   | Launched    | ARB standard of care |
| 10   | eprosartan                   | -5.89   | 0.00  | AGTR1          | 4   | Launched    | ARB standard of care |
| 11   | spp301                       | -5.89   | 0.00  | EDNRA          | 4   | Preclinical | No                   |
| 12   | pyrrolidine-dithiocarbamate  | -5.79   | 0.00  | REL/RELA       | 4   | Preclinical | No                   |
| 14   | caffeic-acid-phenethyl-ester | -5.69   | 0.00  | REL/RELA       | 4   | Preclinical | No                   |
| 16   | F351                         | -5.69   | 0.00  | TGFB1          | 4   | Phase 2     | No                   |
| 18   | triptolide                   | -5.59   | 0.00  | REL/RELA       | 4   | Preclinical | Informal use (TCM)   |
| 19   | bay-11-7082                  | -5.59   | 0.00  | REL/RELA       | 4   | Preclinical | No                   |
| 21   | NF-kB inhibitor II           | -5.50   | 0.00  | REL/RELA       | 4   | Preclinical | No                   |
| 22   | olmesartan                   | -5.50   | 0.00  | AGTR1          | 4   | Launched    | ARB standard of care |
| 25   | FT011                        | -5.17   | 0.00  | TGFB1          | 4   | Preclinical | No                   |
| 35   | bafetinib                    | -3.92   | 0.50  | LYN            | 2   | Phase 2     | No                   |
| 40   | tolimidone                   | -3.79   | 0.75  | LYN            | 2   | Phase 2     | No                   |
| 53   | masitinib                    | -2.43   | 0.75  | LYN            | 2   | Launched    | No                   |
| 55   | rebastinib                   | -2.09   | 0.80  | LYN            | 2   | Phase 1/2   | No                   |
| 56   | bosutinib                    | -2.06   | 0.83  | LYN            | 2   | Launched    | No                   |
| 57   | nintedanib                   | -2.02   | 0.80  | LYN            | 2   | Launched    | No                   |
| 58   | tranilast                    | -1.63   | 0.00  | TGFB1          | 4   | Launched    | No                   |
| 59   | ponatinib                    | -1.59   | 0.50  | LYN            | 2   | Launched    | No                   |
| 61   | dasatinib                    | -0.91   | 0.92  | LYN            | 2   | Launched    | No                   |

Dist. is the observed mean minimum distance from drug targets to genome-wide-association seed proteins. Known drugs for this disease (atrasentan rank 8, fostamatinib rank 188) are excluded here and appear in the Results text; sparsentan (rank 1) is likewise excluded, as its apparent rank reflects post-approval annotation backfill rather than a retrospective signal (see Results). ARB, angiotensin receptor blocker; these are standard-of-care agents for blood pressure and proteinuria control in this disease.

### Genetic association screen

The cross-join of 3,005 gene associations with the Drug Repurposing Hub produced 5,995 drug-gene rows covering 3,236 compounds. Among the top-100 drugs by genetic association score, the LYN locus dominated, with five kinase inhibitors (bafetinib, bosutinib, masitinib, ponatinib, nintedanib) appearing through LYN inhibition. Transforming-growth-factor-beta inhibitors appeared through the TGFB1 locus. Fostamatinib appeared in the top-100 through its SYK target, matching its Phase 2 trial status, confirming that this method also recovers clinically validated compounds.

## GraphSAGE versus relation-aware graph neural network architecture

A graph neural network link predictor was implemented twice on a 35,629-node, 693,524-edge PrimeKG subgraph, first with a relation-agnostic GraphSAGE encoder<sup>10</sup> and then with a relational graph convolutional network<sup>11</sup> that assigns a separate weight matrix per relation type. Both were evaluated under identical leave-one-out cross-validation across the same 15 positive examples (Figure 3). The relation-aware encoder achieved higher recall (13 of 15 known drugs in the top 10% of 7,957 candidates, versus 10 of 15 for the relation-agnostic encoder), and the final 15-candidate shortlist is unchanged between the two encoders, since shortlisting is driven by network proximity and genetic association convergence rather than the graph neural network arm. The two encoders disagree most on sparsentan and atrasentan, the two drugs whose single-split training-set membership motivated the leave-one-out redesign in the first place: sparsentan falls outside the top decile under the relation-agnostic encoder (15.8th percentile) despite ranking well under the relation-aware encoder (2.6th percentile), while atrasentan shows the reverse pattern, ranking more strongly under the relation-agnostic encoder (0.4th percentile) than the relation-aware one (3.1th percentile).

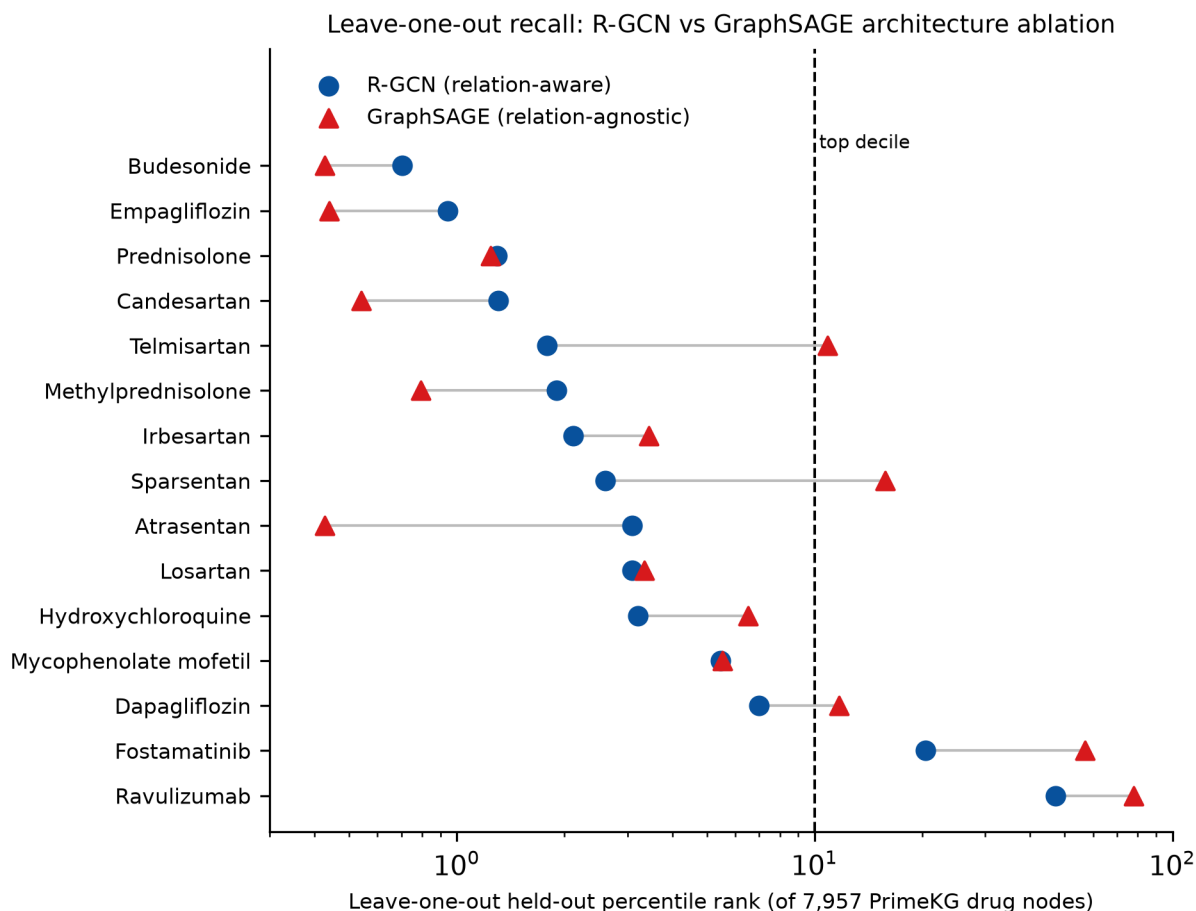


**Figure 3.** Leave-one-out recall: relation-aware (R-GCN) versus relation-agnostic (GraphSAGE) architecture ablation. For each of the 15 known drugs resolvable as PrimeKG drug nodes, a fresh model was trained with that drug excluded from the positive supervision set, and its percentile rank among all 7,957 PrimeKG drug nodes was recorded, giving every drug exactly one genuinely blind evaluation per encoder. Circles, R-GCN; triangles, GraphSAGE. Dashed line, top decile. 13 of 15 known drugs recover in the top decile under R-GCN, versus 10 of 15 under GraphSAGE.

## Consensus shortlist and clinical feasibility

Intersecting the network proximity and genetic association top-100 lists produced 17 candidates, of which two non-therapeutic compounds were removed, leaving 15 final candidates with no trial record for this disease (Table 2). These

divide into three mechanism clusters (Figure 4). A LYN kinase cluster contains nine candidates across a range of clinical stages, one of which, nintedanib, an approved inhibitor of PDGFR, FGFR, VEGFR, and LYN developed for idiopathic pulmonary fibrosis<sup>13</sup>, addresses both the LYN and TGF- $\beta$  axes simultaneously and received the highest clinical priority score. A transforming-growth-factor-beta cluster contains five candidates, including a Phase 2 compound, F351<sup>12</sup>, whose target is itself a genome-wide-association seed protein with zero network distance to the disease module. A single REL/NF-kB candidate, ginsenoside-RE, completes the set through the same NF-kB axis recovered independently in the broader network proximity screen (Table 1). Graph neural network and transcriptomic reversal scores are available for a subset of these 15: eight candidates are absent from PrimeKG's drug vocabulary and receive no graph neural network score, and eight (a different but overlapping set) are absent from LINCS L1000 and receive no transcriptomic reversal score (Figure 4; Table 3); these are coverage gaps in the respective databases rather than negative results.

**Table 2.** Final 15 candidates after consensus shortlist and clinical feasibility filter.

| Rank | Drug           | Phase       | Hit | Target | NP rank | NP z  | OT rank | eGFR          | Neph     | BP                 |
|------|----------------|-------------|-----|--------|---------|-------|---------|---------------|----------|--------------------|
| 1    | nintedanib     | Launched    | 2+4 | LYN    | 57      | -2.02 | 13      | monitor       | low      | none               |
| 2    | bosutinib      | Launched    | 2   | LYN    | 56      | -2.06 | 7       | reduce<br><30 | low      | low                |
| 3    | masitinib      | Launched    | 2   | LYN    | 53      | -2.43 | 14      | monitor       | low      | low                |
| 4    | tranilast      | Launched    | 4   | TGFB1  | 58      | -1.63 | 22      | monitor       | low      | none               |
| 5    | adaprev        | Phase 3     | 4   | TGFB1  | 5       | -6.52 | 23      | monitor       | low      | none               |
| 6    | dasatinib      | Launched    | 2   | LYN    | 61      | -0.91 | 15      | monitor       | low      | moderate           |
| 7    | F351           | Phase 2     | 4   | TGFB1  | 16      | -5.69 | 20      | monitor       | low      | none               |
| 8    | bafetinib      | Phase 2     | 2   | LYN    | 35      | -3.92 | 12      | monitor       | low      | low                |
| 9    | tolimidone     | Phase 2     | 2   | LYN    | 40      | -3.79 | 11      | none          | none     | none               |
| 10   | rebastinib     | Phase 1/2   | 2   | LYN    | 55      | -2.09 | 10      | monitor       | low      | low                |
| 11   | ginsenoside-RE | Phase 1     | 4   | REL    | 30      | -4.26 | 16      | none          | none     | none               |
| 12   | ponatinib      | Launched    | 2   | LYN    | 59      | -1.59 | 9       | monitor       | moderate | high <sup>19</sup> |
| 13   | D-4476         | Preclinical | 4   | TGFB1  | 6       | -6.38 | 21      | -             | -        | -                  |
| 14   | FT011          | Preclinical | 4   | TGFB1  | 27      | -5.17 | 19      | -             | -        | -                  |
| 15   | ON123300       | Preclinical | 2   | LYN    | 54      | -2.20 | 8       | -             | -        | -                  |

NP, network proximity. OT, Open Targets genetic association rank (of 3,005). eGFR, dose-adjustment flag. Neph, nephrotoxicity risk. BP, hypertension risk. Ponatinib's BP risk (67% incidence at standard oncology doses<sup>19</sup>) requires specialist cardiovascular review before any investigation in this disease. Bosutinib requires dose reduction to 200 mg/day at eGFR below 30 mL/min/1.73 m<sup>2</sup>, per the prescribing information<sup>20</sup>. Safety annotations are derived from FDA prescribing information; see Discussion for caveats.

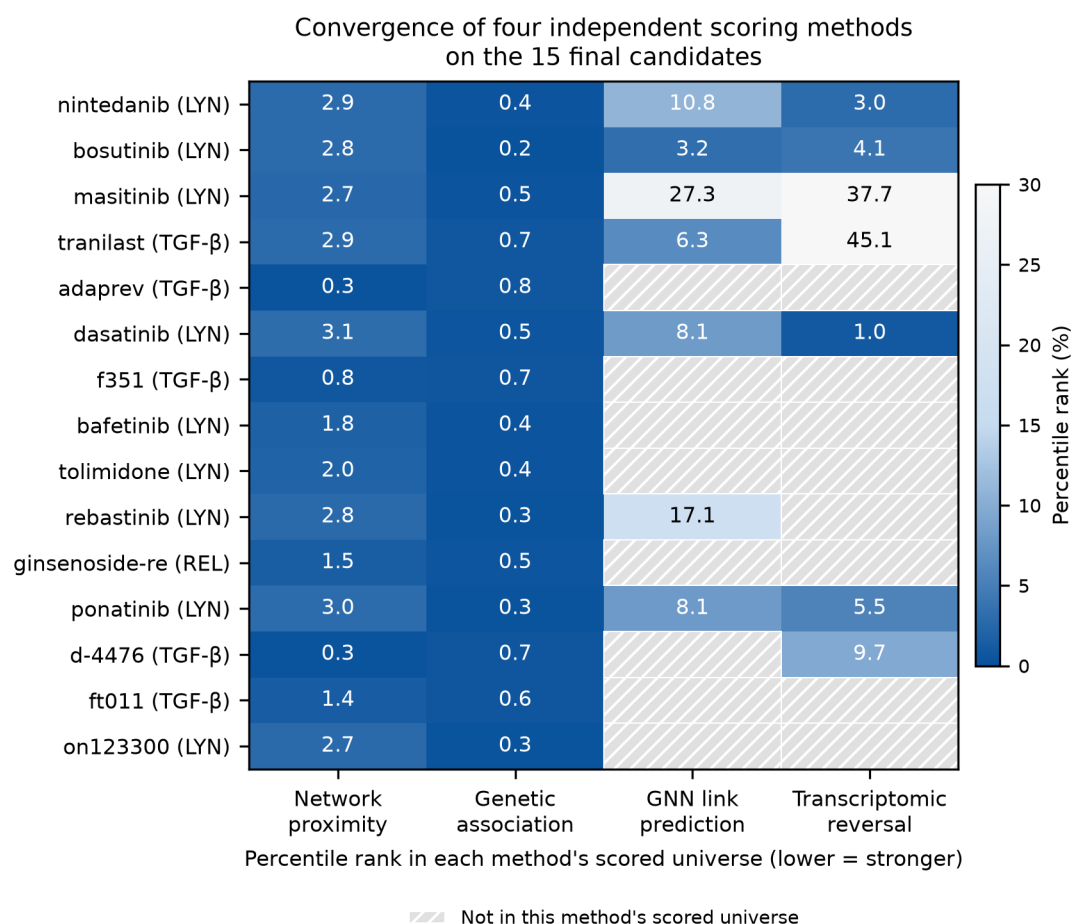


**Figure 4.** Convergence of four independent scoring methods on the 15 final candidates. Heatmap cells show each candidate's percentile rank within each method's own scored universe (network proximity, 1,989 compounds; genetic association, 3,005 genes; graph neural network link prediction, 7,957 drug nodes; transcriptomic reversal, 3,344 drugs); darker shading indicates a stronger (lower percentile) rank. Network proximity and genetic association are strong for all 15 candidates by construction, since the consensus shortlist is defined as their intersection. Graph neural network and transcriptomic reversal are independent of that selection step and show genuine variation: most LYN-cluster candidates rank in the top 10% on both, while others are weaker on one axis or absent from that method's scored universe entirely (hatched cells), most often because the compound is not represented in the underlying knowledge graph or expression-perturbation database rather than because it scored poorly.

### Dossier evidence for top candidates

Evidence dossiers were generated for the top-10 candidates by clinical priority score. A 2026 case report<sup>14</sup> describes IgA-nephropathy-pattern glomerulonephritis on renal biopsy in a patient treated with nintedanib for idiopathic pulmonary fibrosis, indicating that the drug reaches renal tissue and engages IgA-nephropathy-relevant biology in humans; therapeutic benefit at a disease-specific dose remains unestablished.

### Transcriptomic reversal analysis

A Connectivity Map analysis against a glomerular differential expression signature (20 disease biopsies versus 22 healthy controls<sup>15</sup>) scored 3,344 drugs from the LINCS L1000 compendium via Enrichr<sup>16</sup> (Table 3). Five of the seven consensus candidates present in LINCS ranked in the top 10% for transcriptomic reversal, including four of the five LYN-cluster candidates: dasatinib, nintedanib, bosutinib, and ponatinib, between the 1.0th and 5.5th percentiles. This provides a fourth

independent validation signal from a completely different data type, converging with network proximity, genetic association, and graph neural network link prediction on the LYN kinase cluster.

**Table 3.** Transcriptomic reversal scores for consensus candidates present in LINCS L1000.

| Candidate  | Mechanism cluster   | CMAP rank (of 3,344) | Percentile | Connectivity score | Top 10% |
|------------|---------------------|----------------------|------------|--------------------|---------|
| Dasatinib  | LYN kinase          | 32                   | 1.0        | 334.6              | Yes     |
| Nintedanib | LYN +<br>PDGFR/FGFR | 99                   | 3.0        | 179.2              | Yes     |
| Bosutinib  | LYN kinase          | 136                  | 4.1        | 131.3              | Yes     |
| Ponatinib  | LYN kinase          | 185                  | 5.5        | 99.1               | Yes     |
| D-4476     | TGF-beta/CK1a       | 323                  | 9.7        | 48.4               | Yes     |
| Masitinib  | LYN kinase          | 1,262                | 37.7       | 13.4               | No      |
| Tranilast  | TGF-beta            | 1,508                | 45.1       | 11.1               | No      |

Disease signature: 200 up-regulated and 200 down-regulated genes from glomerular biopsies of this disease versus healthy controls<sup>15</sup>. Connectivity score is the geometric mean of Enrichr combined scores across directional LINCS L1000 queries<sup>16</sup>. Eight candidates absent from LINCS (adaprev, F351, bafetinib, FT011, on123300, tolimidone, rebastinib, ginsenoside-RE) were not scored; absence is a database coverage gap, not a negative result.

### Protein-ligand structure prediction

Protein-ligand co-folding was performed for the five top-ranked consensus candidates using an open-weight structure prediction model<sup>17</sup>. Four of five drug-protein pairs scored a ligand interface predicted TM-score of 0.85 or above (mean of three diffusion samples; Supplementary Table 1), with the four LYN kinase inhibitors clustering between 0.93 and 0.98. Contact residue analysis confirmed that LYN inhibitors consistently engage the canonical ATP-binding pocket across diffusion samples. The transforming-growth-factor-beta-1 candidate scored lower (0.77) with no contact overlap across samples, consistent with this target being a secreted cytokine without a canonical small-molecule pocket rather than a kinase with a defined active site.

## Discussion

This study presents a systematic, four-method computational repurposing screen for IgA nephropathy, a disease whose five recent drug approvals have not been matched by any computational drug discovery effort. Fifteen candidates are nominated across three mechanism clusters, none with a prior trial record in this disease.

The strongest candidate is nintedanib, an approved inhibitor of LYN, PDGFR, FGFR, and VEGFR developed for idiopathic pulmonary fibrosis<sup>13</sup>. In this disease, PDGFR-beta drives mesangial cell proliferation, one of the primary pathological findings on biopsy, while LYN sits at the apex of the B-cell receptor cascade that drives production of the pathogenic autoantibodies against galactose-deficient IgA1. No current approved therapy addresses both axes simultaneously. Four independent lines of evidence support this candidate: a transcriptomic reversal signal at the 3.0th percentile of 3,344 drugs, using no network or genetic data; a 2026 case report<sup>14</sup> of IgA-nephropathy-pattern glomerulonephritis in a treated patient, which most plausibly reflects unmasking of pre-existing subclinical disease through on-target pharmacology rather than de novo induction, since the histological pattern matches established rather than drug-induced nephropathy; a demonstrated safety profile at reduced eGFR with no dose adjustment required<sup>13</sup>; and a structural co-folding confidence of 0.97 with consistently predicted ATP-pocket contacts. The priority experiment is an in vitro study in human mesangial cells and a patient-derived model, followed by testing in a spontaneous rodent model of the disease, before any clinical trial design.

The fibrosis component of this disease receives the least therapeutic attention among the five approved agents: none directly inhibits transforming growth factor beta signalling. F351<sup>12</sup>, a Phase 2 compound in the consensus shortlist, targets this locus

directly, with zero network distance to the disease module, since its target is itself a genome-wide-association seed protein. A second, launched compound, tranilast, with decades of off-label use in fibrotic conditions and documented renoprotective effects in animal models, ranked in the consensus shortlist independently. At a 10% network-proximity threshold with six- to seven-fold enrichment over random, and restricting the retrospective recall claim to the two drugs whose targets are genuinely annotation-safe at the relevant data vintage, recall is 2 of 2 for mechanism-captured, annotation-safe cases. Estimated precision for the 15-drug shortlist, based on published benchmarks for consensus computational repurposing pipelines<sup>18</sup>, is in the range of 20-40%, implying 3-6 expected true positives and 9-12 expected false positives from a 500-fold reduction of the original compound library; this is not a failure of the method but reflects a mechanism-grounded prior producing a feasible experimental set rather than a guaranteed hit list. Two further approved drugs, iptacopan and sibeprenlimab, are absent from the Drug Repurposing Hub and could not be scored by this method, but one drug’s target independently reached the top 5% of genetic-association evidence roughly seven years before its own approval, using a gene-level signal with no drug-annotation dependency and hence no equivalent backfill risk; a dedicated dated-snapshot audit of this pipeline, in preparation, reports that result in full.

Fostamatinib, a SYK inhibitor one hop downstream of LYN, is in Phase 2 trials for this disease and ranked at the 9.5th percentile in network proximity, matching its known one-hop distance from the LYN seed. Direct LYN inhibitors have shorter network distances and correspondingly stronger proximity scores, providing a mechanistic gradient consistent with the hypothesis that targeting the upstream kinase may produce broader pathway suppression than targeting SYK alone; whether this translates into superior clinical efficacy is an empirical question this pipeline nominates rather than answers.

Several limitations qualify these results. The graph neural network’s positive training set is small (15 examples), and while leave-one-out cross-validation avoids the circularity of a single random split, the architecture remains a third confirming signal rather than a primary finding, pending positive-set augmentation and pretrained protein-language-model embeddings. The network proximity background is the 953-protein disease module itself rather than the full interactome, which is conservative for the reported associations but may under-rank drugs acting through mechanisms distal from the genetic loci. Clinical safety annotations are derived from regulatory label summaries rather than primary pharmacovigilance data and require independent verification before any clinical use. Structural predictions reflect model self-consistency rather than experimental binding affinity. This pipeline operates entirely on public data with no experimental component, and all candidates require in vitro and in vivo validation before any clinical trial is designed.

To our knowledge, no prior published study has applied graph neural network drug repurposing, knowledge graph embedding, network proximity analysis, or multi-omics target prioritisation to this disease, based on supplementary literature searches that returned no prior computational repurposing studies. This gap is not unique to this disease: nephrology as a field receives an estimated 0.6% of artificial-intelligence drug discovery publications despite accounting for 11% of global mortality from chronic disease<sup>3</sup>. The pipeline presented here is transferable to other glomerular diseases by substituting the genome-wide-association seed list and disease node identifier, and its central methodological contribution, a leakage-aware validation protocol that separates genuine retrospective signal from post-approval annotation backfill, is intended to generalise beyond this disease and this pipeline.

## Conclusions

A leakage-aware protocol changes what a retrospective recall claim can support. Applied to IgA nephropathy, it removed the pipeline’s highest-ranked apparent success as annotation backfill, kept two annotation-safe recoveries in the top decile, and left a fifteen-candidate shortlist on which four independent methods converge for LYN kinase inhibition. The protocol needs only a dated copy of each reference file and a held-out evaluation design, so it transfers to other repurposing pipelines and other diseases.

## Data and Software Availability

All primary data sources are open access: PrimeKG (Harvard Dataverse)<sup>5</sup>, the Drug Repurposing Hub (Broad Institute; drug information files dated 24 March 2020 and 18 August 2025)<sup>6</sup>, STRING v12<sup>8</sup>, the Open Targets Platform<sup>7</sup>, GEO dataset

GSE93798<sup>15</sup>, and LINCS L1000 via Enrichr<sup>16</sup>. Analysis scripts (Python 3), processed data, ranked outputs for every method, candidate dossiers, and figure-generation code are archived at Zenodo (<https://doi.org/10.5281/zenodo.23098959>).

## Supporting Information

Supplementary Table 1: protein-ligand co-folding confidence and contact residues for the five top-ranked candidates (PDF).

## Author Information

Corresponding author: Hanif Rahman, Independent Researcher, London, UK; [hanif@hanifrahman.com](mailto:hanif@hanifrahman.com); ORCID <https://orcid.org/0009-0004-4270-4222>.

Author contributions: H.R. conceived and designed the study, implemented all pipeline scripts, performed all analyses, interpreted the results, and wrote the manuscript.

Funding: no funding was received for this work.

Notes: the author has filed two UK patent applications relevant to this work: GB2613032.8 (filed 4 June 2026), covering second medical use claims for candidates including nintedanib, dasatinib, bosutinib, and F351 in IgA nephropathy, arising from the pipeline described here; and GB2613961.8 (filed 16 June 2026), covering a target-prioritisation algorithm applied in a companion study and not the subject of this manuscript. Neither filing restricts publication of the methods or results, and no data or analysis has been withheld or altered on account of either.

## Acknowledgments

The author thanks the teams behind the Drug Repurposing Hub (Broad Institute), PrimeKG (Harvard Dataverse), the Open Targets Platform, and the STRING database. A large language model (Claude, Anthropic) assisted with drafting manuscript prose and debugging analysis and figure-generation code. All code was executed and all results independently verified by the author, who takes full responsibility for the content, accuracy, and conclusions of this work. No generative AI tool is credited as an author, and no figure was generated or substantively modified by AI image generation.

## References

1. Kiryluk, K.; et al. Genome-Wide Association Analyses Define Pathogenic Signaling Pathways and Prioritize Drug Targets for IgA Nephropathy. *Nat. Genet.* **2023**, *55*, 1091-1105.
2. U.S. Food and Drug Administration. Drugs@FDA: FDA-Approved Drugs. <https://www.accessdata.fda.gov/scripts/cder/daf/> (accessed 2025).
3. Zhuang, K.; Wang, W.; Xu, C.; Guo, X.; Ren, X.; Liang, Y.; Duan, Z.; Song, Y.; Zhang, Y.; Cai, G. Machine Learning-Based Diagnosis and Prognosis of IgAN: A Systematic Review and Meta-Analysis. *Heliyon* **2024**, *10*, e33090.
4. Guney, E.; Menche, J.; Vidal, M.; Barabási, A.-L. Network-Based in Silico Drug Efficacy Screening. *Nat. Commun.* **2016**, *7*, 10331.
5. Chandak, P.; Huang, K.; Zitnik, M. Building a Knowledge Graph to Enable Precision Medicine. *Sci. Data* **2023**, *10*, 67.
6. Corsello, S. M.; et al. The Drug Repurposing Hub: A Next-Generation Drug Library and Information Resource. *Nat. Med.* **2017**, *23*, 405-408.
7. Ochoa, D.; et al. The Next-Generation Open Targets Platform: Reimagined, Redesigned, Rebuilt. *Nucleic Acids Res.* **2023**, *51*, D1353-D1359.
8. Szklarczyk, D.; et al. The STRING Database in 2023: Protein-Protein Association Networks and Functional Enrichment Analyses for Any of 14,093 Organisms. *Nucleic Acids Res.* **2023**, *51*, D638-D646.
9. Wang, Z.; Yu, C.; Zhou, L. N.; Chen, X. Effects of Tripterygium wilfordii Induction Therapy to IgA Nephropathy Patients with Heavy Proteinuria. *Biol. Pharm. Bull.* **2017**, *40*, 1833-1838.

10. Hamilton, W. L.; Ying, R.; Leskovec, J. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems* 30; 2017.
11. Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; van den Berg, R.; Titov, I.; Welling, M. Modeling Relational Data with Graph Convolutional Networks. In *The Semantic Web: ESWC 2018*; Springer, 2018; pp 593-607.
12. Cai, X.; et al. Hydronidone for the Treatment of Liver Fibrosis Related to Chronic Hepatitis B: A Phase 2 Randomized Controlled Trial. *Clin. Gastroenterol. Hepatol.* **2023**, *21*, 1893-1901.
13. Richeldi, L.; et al. Efficacy and Safety of Nintedanib in Idiopathic Pulmonary Fibrosis. *N. Engl. J. Med.* **2014**, *370*, 2071-2082.
14. Kunisho, Y.; et al. A Case of Membranoproliferative Glomerulonephritis-Type IgA Nephropathy Modified by Nintedanib. *CEN Case Rep.* **2026**, *15*, 44. Correction: *CEN Case Rep.* **2026**, *15*, 63.
15. Liu, P.; et al. Transcriptomic and Proteomic Profiling Provides Insight into Mesangial Cell Function in IgA Nephropathy. *J. Am. Soc. Nephrol.* **2017**, *28*, 2961-2972.
16. Kuleshov, M. V.; et al. Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update. *Nucleic Acids Res.* **2016**, *44*, W90-W97.
17. Passaro, S.; et al. Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. *bioRxiv* **2025**, DOI: 10.1101/2025.06.14.659707.
18. Pushpakom, S.; et al. Drug Repurposing: Progress, Challenges and Recommendations. *Nat. Rev. Drug Discovery* **2019**, *18*, 41-58.
19. Cortes, J. E.; et al. A Phase 2 Trial of Ponatinib in Philadelphia Chromosome-Positive Leukemias. *N. Engl. J. Med.* **2013**, *369*, 1783-1796.
20. Pfizer Inc. BOSULIF (bosutinib) Prescribing Information. [https://www.accessdata.fda.gov/drugsatfda\\_docs/label/2024/203341s026lbl.pdf](https://www.accessdata.fda.gov/drugsatfda_docs/label/2024/203341s026lbl.pdf) (accessed 2025).

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [p01preprintsupplementaryv1.pdf](#)