ARIA
APPLIED RESEARCH IN ACTION

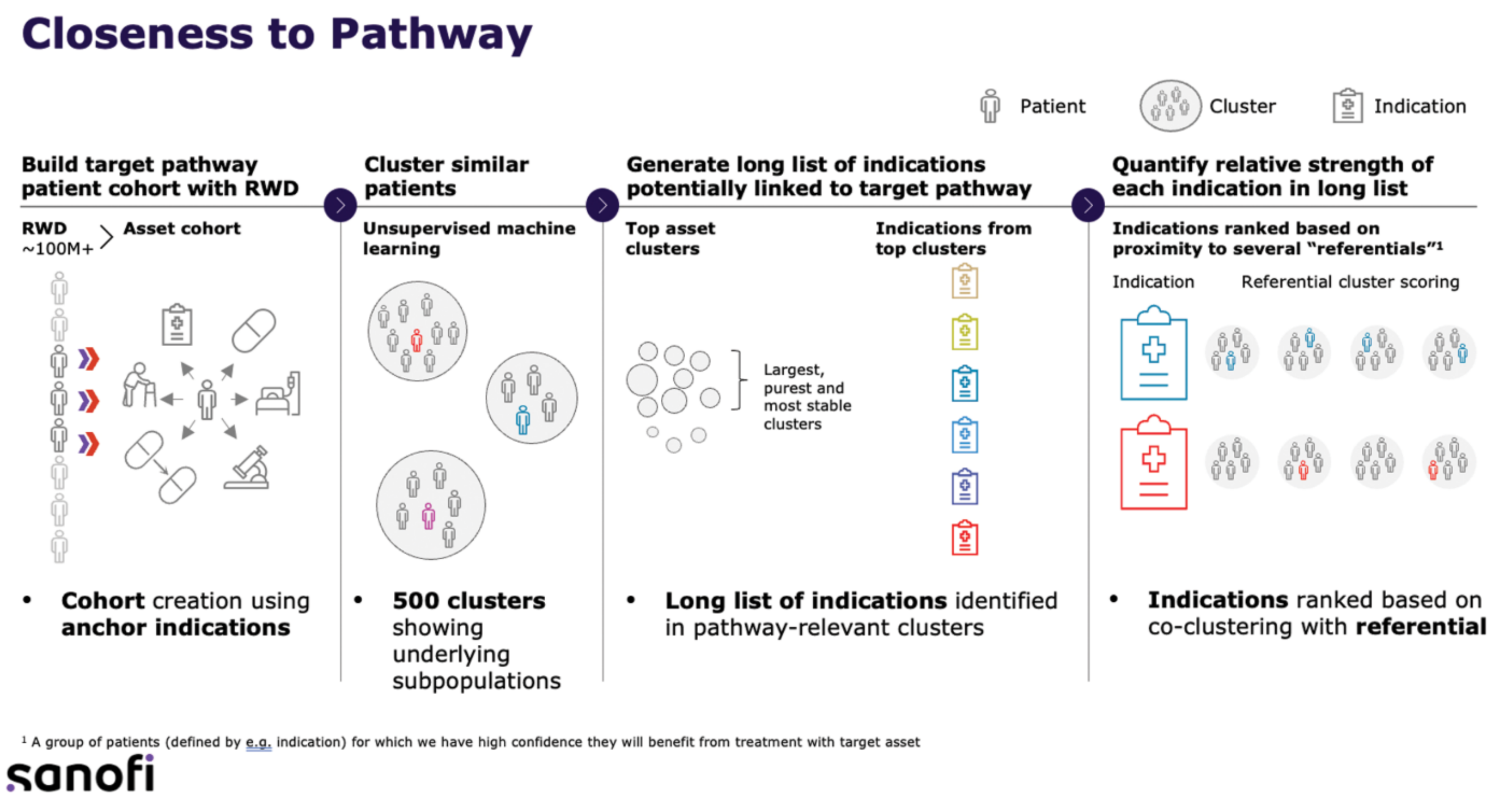
Closeness to Pathway: Clustering Patient Electronic Medical Records for Drug Repurposing

Evaluating and comparing drug repositioning approaches that use unsupervised learning and patient data from electronic medical records against supervised learning with known drug-disease relationship.

Chaorui Zhang

Anna Goldenberg
ACADEMIC SUPERVISOR

Brandon Rufino
INDUSTRY SUPERVISOR



Closeness to Pathway

Patient · Cluster · Indication

Build target pathway patient cohort with RWD
RWD ~100M+ > Asset cohort
- **Cohort** creation using **anchor indications**

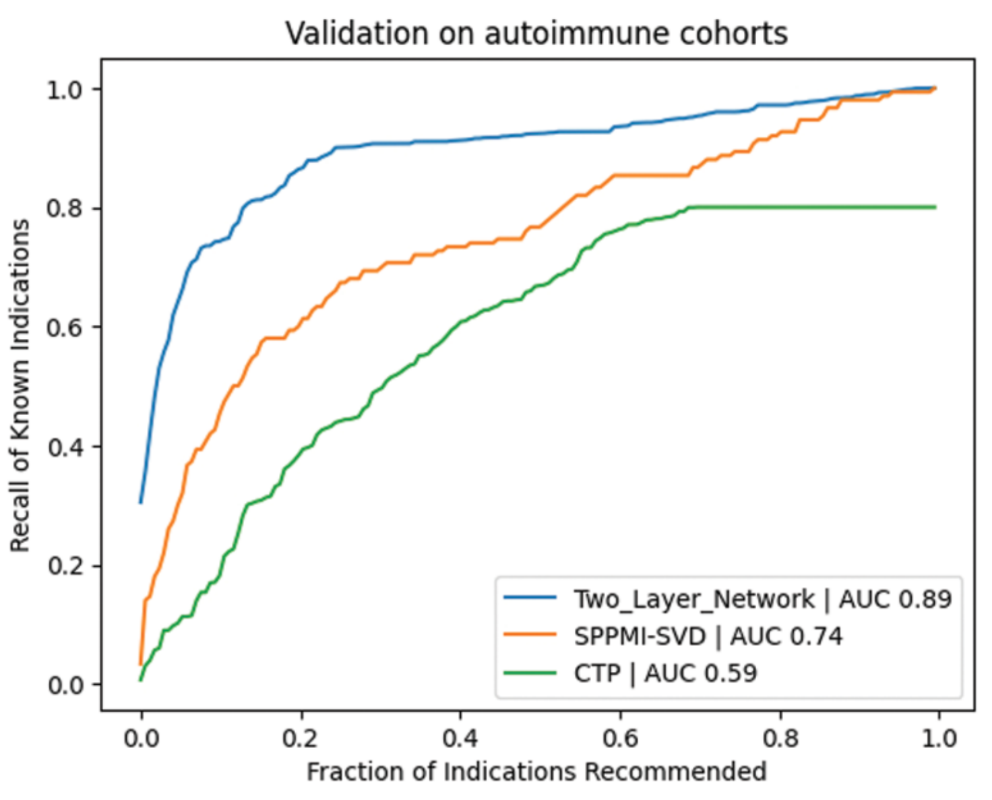
Cluster similar patients
Unsupervised machine learning
- **500 clusters** showing underlying subpopulations

Generate long list of indications potentially linked to target pathway
Top asset clusters · Largest, purest and most stable clusters · Indications from top clusters
- **Long list of indications** identified in pathway-relevant clusters

Quantify relative strength of each indication in long list
Indications ranked based on proximity to several "referentials"[1]
Indication · Referential cluster scoring
- **Indications** ranked based on co-clustering with **referential**

[1] A group of patients (defined by e.g. indication) for which we have high confidence they will benefit from treatment with target asset

sanofi



PROJECT SUMMARY

Drug repositioning involves identifying unaddressed diseases for existing and new therapeutics, helping subject matter experts narrow their options from hundreds of disease candidates to a handful. Computational approaches involve the use of similarity measures from different modalities (e.g., genetics, side-effects, chemical structures) to infer on new drug-disease relationships based on existing knowledge. These approaches, however, may suffer in inferring diseases for first-in-class therapeutics because a lack of drug-disease relationship is treated implicitly as a negative one. Electronic medical records (EMRs) capture information from a wide range of patients with varying demographics, co-morbidities, and genetic backgrounds. This diversity can help identify unexpected drug-disease relationships that might not be evident from known similarities alone. In this work, we present computational drug repositioning approaches using real-world data and unsupervised learning. First, we designed an algorithm called 'Closeness to Pathway (CTP)' based on patients' EMRs and claims. CTP aims to provide diseases to pursue based on the hypothesis that phenotypical similarity is a proxy for underlying disease mechanism. Given a cohort of patients with their medical records in a therapeutic area, CTP derives patient features, clusters patients with similar phenotypes and produces a ranked list of diseases that are most prevalent in clusters with the target pathway. Second, we adopted an approach called 'shifted pairwise positive mutual information - singular value decomposition' (SPPMI-SVD) which incorporates locality of EMR events for deriving feature embeddings to use downstream for disease embedding comparison. Lastly, we provided a benchmark following traditional approaches trained on existing drug-disease relationships called 'two-layer heterogeneous network'. We evaluated the performance of all three algorithms by constructing validation sets in immunology, inflammatory and oncology cohorts using the Unified Medical Language System (UMLS). We found the two-layer heterogenous network outperformed our unsupervised approaches in overall accuracy. However, CTP and SPPMI-SVD were able to correctly identify diseases where limited evidence of drug class to therapeutic area was observed whereas the two-layer heterogeneous network struggled. This work presents opportunities to integrate real-world data with popular computational approaches in drug repurposing.

REFERENCES

Wen J, Zhang X, Rush E, Panickan VA, Li X, Cai T, Zhou D, Ho YL, Costa L, Begoli E, Hong C, Gaziano JM, Cho K, Lu J, Liao KP, Zitnik M, Cai T. Multimodal representation learning for predicting molecule-disease relations. Bioinformatics. 2023 Feb 3;39(2):btad085. doi: 10.1093/bioinformatics/btad085. PMID: 36805623; PMCID: PMC9940625.

sanofi

Computer Science
UNIVERSITY OF TORONTO

Master of Science in Applied Computing