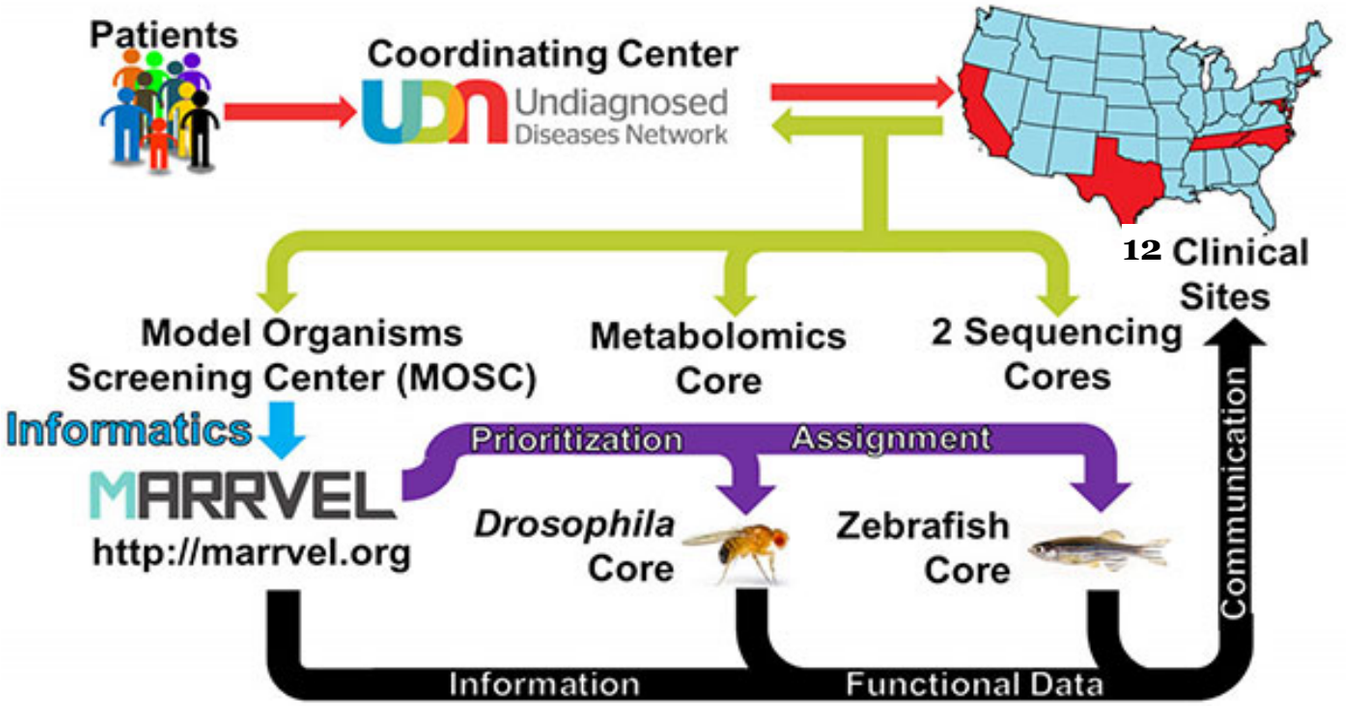
Attentive Multiview Text Representation for Differential Diagnosis

Hadi Amiri, Mitra Mohatarami, Isaac S. Kohane
UMASS, Lowell; MIT; Harvard
hadi_amiri@uml.edu, mitram@mit.edu, isaac_kohane@harvard.edu

HARVARD
MEDICAL SCHOOL

Motivation

- Undiagnosed Diseases Network (UDN)
  - Provide diagnosis for undiagnosed patients
    - Reducing health disparities
    - Prioritizing genetic variants

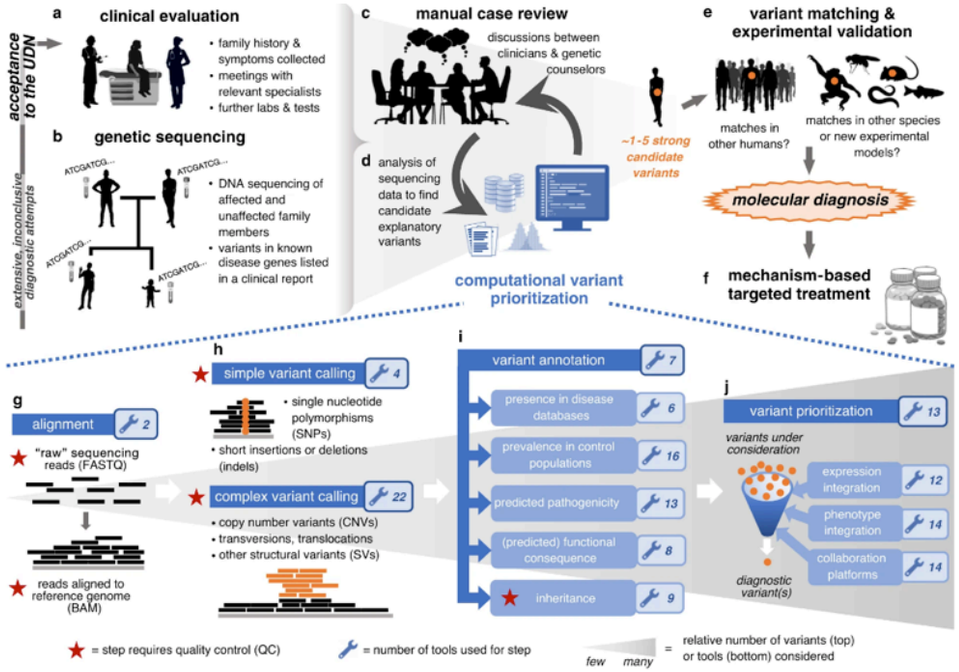


Clinical Sites

- Bethesda, MD (NIH)
- Boston, MA (Harvard)
- Durham, NC (Duke and Columbia)
- Houston, TX (Baylor)
- Los Angeles, CA (UCLA)
- Miami, FL (Miami School of Medicine)
- Nashville, TN (Vanderbilt)
- Philadelphia, PA (Children's Hospital)
- Salt Lake City, UT (Utah)
- Seattle, WA (Washington)
- Stanford, CA (Stanford)
- St. Louis, MO (Washington St. Louis)

Motivation

- Undiagnosed Diseases Network (UDN)
  - Patient demographic data
  - Symptoms & exposure info
  - Referral letters
  - Sequencing data
  - Sequencing reports
  - Resource materials about rare diseases
  - Etc.



Source: Kobren et al. (Genet Med. 2021). Commonalities across computational workflows for uncovering explanatory variants in undiagnosed cases.

Effectively integrate different types of patient data to better find matching disorders

Motivation

- Undiagnosed Diseases Network (UDN)
  - Patient demographic data
  - Symptoms & exposure info
  - Referral letters
  - Sequencing data
  - Sequencing reports
  - Resource materials about rare diseases
  - Etc.

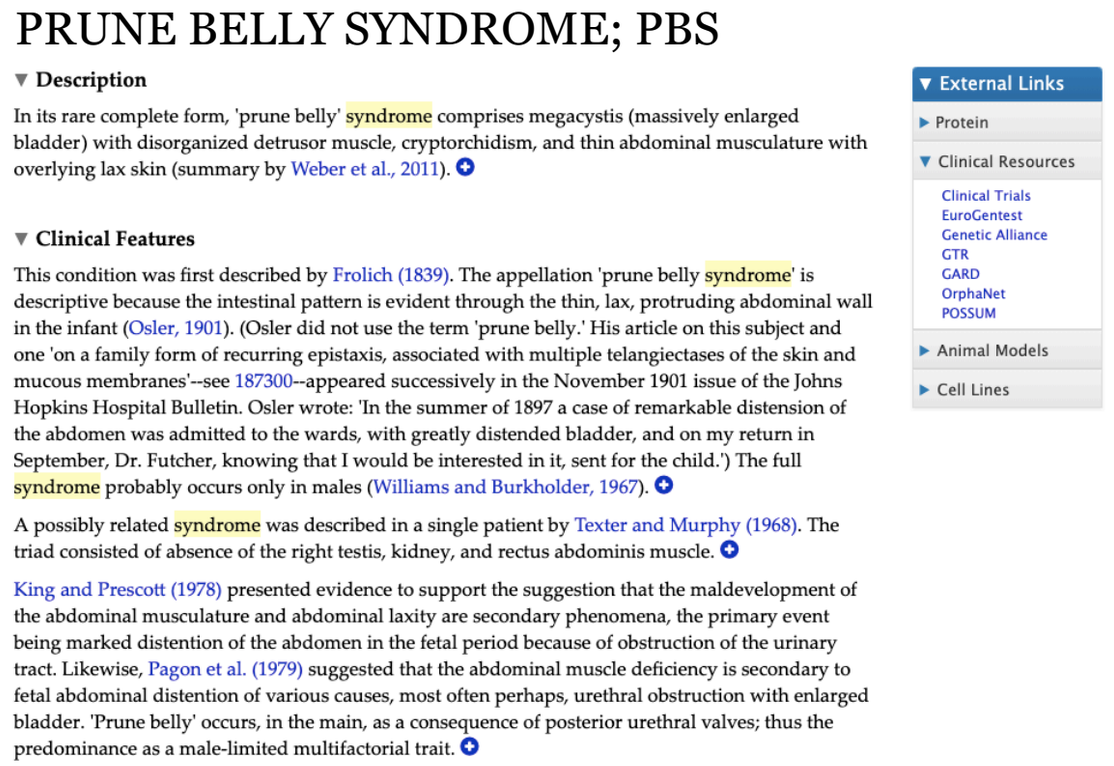
PHYSICIAN-PROVIDED TEXT

- Applicant's medical problems
- Date when symptoms were first noticed
- Previous diagnoses
- History of evaluations and testing
- History of treatments and medications
- Current medications
- Family history
- Diagnostic impressions
- For pediatric patients, prenatal and birth history

Effectively integrate different types of patient data to better find matching disorders

Motivation

- Undiagnosed Diseases Network (UDN)
  - Patient demographic data
  - Symptoms & exposure info
  - Referral letters
  - Sequencing data
  - Sequencing reports
  - Resource materials about rare diseases
  - Etc.



PRUNE BELLY SYNDROME; PBS

▼ **Description**

In its rare complete form, 'prune belly' syndrome comprises megacystis (massively enlarged bladder) with disorganized detrusor muscle, cryptorchidism, and thin abdominal musculature with overlying lax skin (summary by Weber et al., 2011).

▼ **Clinical Features**

This condition was first described by Frolich (1839). The appellation 'prune belly syndrome' is descriptive because the intestinal pattern is evident through the thin, lax, protruding abdominal wall in the infant (Osler, 1901). (Osler did not use the term 'prune belly.' His article on this subject and one 'on a family form of recurring epistaxis, associated with multiple telangiectases of the skin and mucous membranes'--see 187300--appeared successively in the November 1901 issue of the Johns Hopkins Hospital Bulletin. Osler wrote: 'In the summer of 1897 a case of remarkable distension of the abdomen was admitted to the wards, with greatly distended bladder, and on my return in September, Dr. Futcher, knowing that I would be interested in it, sent for the child.') The full syndrome probably occurs only in males (Williams and Burkholder, 1967).

A possibly related syndrome was described in a single patient by Texter and Murphy (1968). The triad consisted of absence of the right testis, kidney, and rectus abdominis muscle.

King and Prescott (1978) presented evidence to support the suggestion that the maldevelopment of the abdominal musculature and abdominal laxity are secondary phenomena, the primary event being marked distention of the abdomen in the fetal period because of obstruction of the urinary tract. Likewise, Pagon et al. (1979) suggested that the abdominal muscle deficiency is secondary to fetal abdominal distention of various causes, most often perhaps, urethral obstruction with enlarged bladder. 'Prune belly' occurs, in the main, as a consequence of posterior urethral valves; thus the predominance as a male-limited multifactorial trait.

▼ External Links
- ▶ Protein
- ▼ Clinical Resources
  - Clinical Trials
  - EuroGentest
  - Genetic Alliance
  - GTR
  - GARD
  - OrphaNet
  - POSSUM
- ▶ Animal Models
- ▶ Cell Lines

www.omim.org

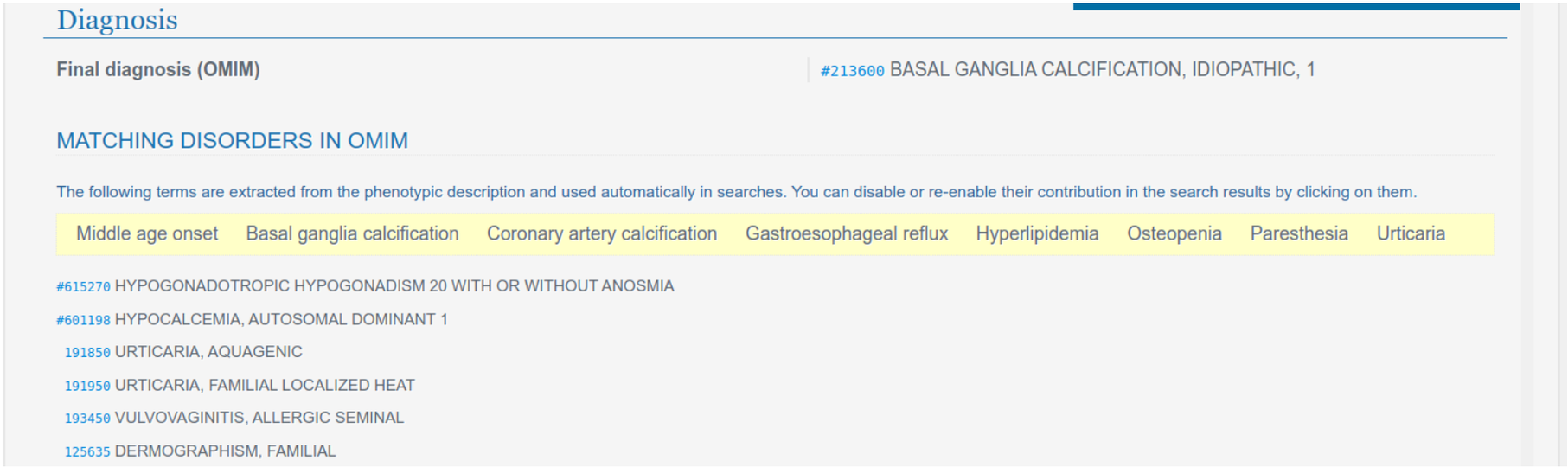
Effectively integrate different types of patient data to better find matching disorders

Ranking Problem

Finding the Zebra!

PhenoTips

- Uses the Human Phenotype Ontology and Orphanet
- Ontology-based similarity to phenotypic descriptions of patients
- Advanced statistical modeling in differentiating candidate diseases
- Accounts for disorder frequencies in the general population
- Supports negation
  - e.g., symptoms not observed in the patient.



Girdea M., et al. Phenotips: Patient phenotyp-ing software for clinical and research use. Humanmutation, 34(8):1057–1065. 2013.

Ranking Problem

- Let ***q*** indicate *representation* of a specific type of **patient** data
- Let ***d*** indicate *representation* of a specific type of **disease** data
- Let (***q***', ***d***'), (***q***'', ***d***''), … denote ***different views*** of the same patient/disease.

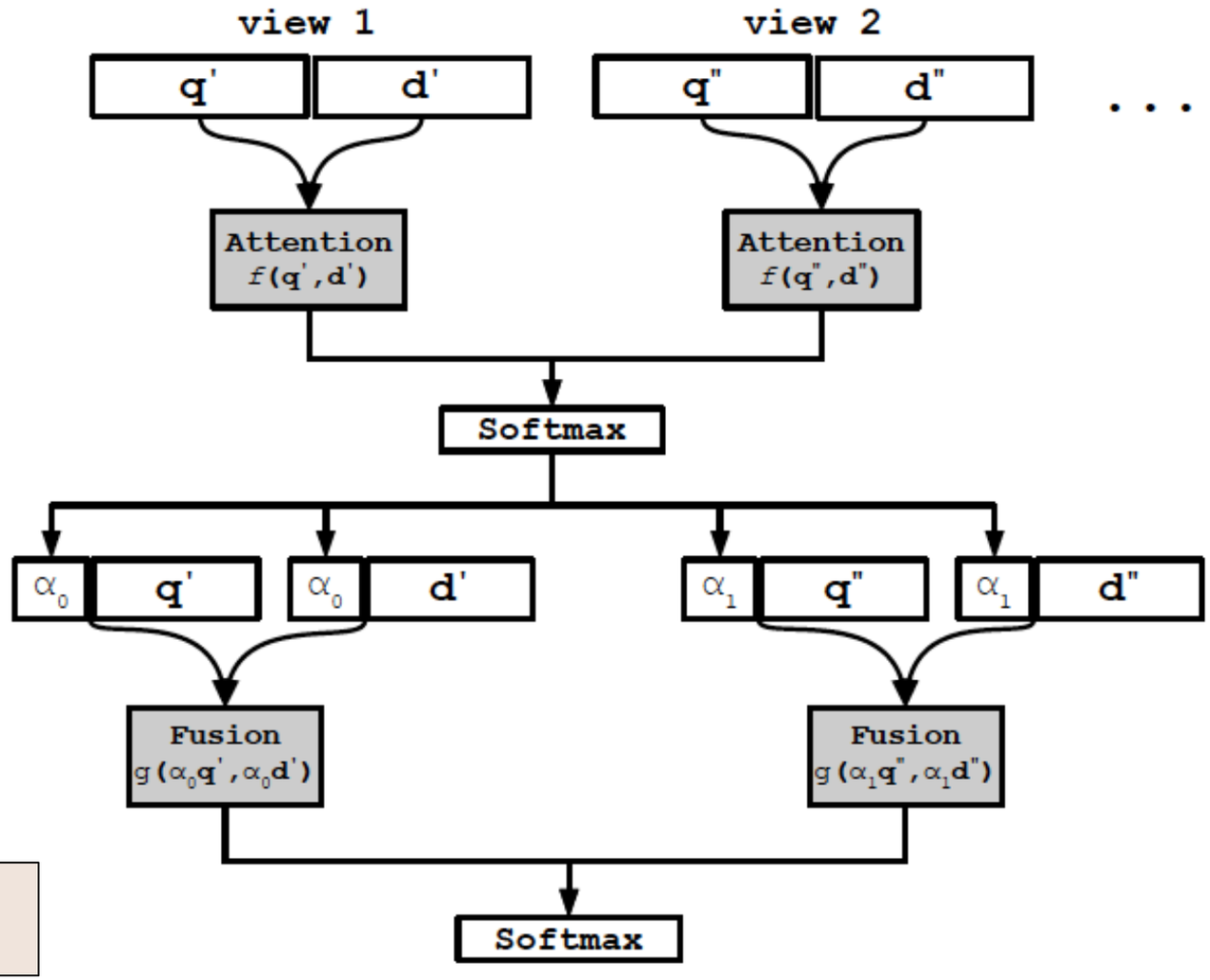
- Task of interest
  - Determine a ***relevance score*** between each given ***q*** and ***d***.

- Solution
  - Effectively prioritize and combine representations through ***Attention*** and ***Fusion*** neural sub-networks.

Ranking Problem

- Attentive Multiview Neural Model (AMNM)
  - $f(.)$ indicates attention network
  - $g(.)$ indicates fusion network
  - a_i indicates the attentive weight of the i^{th} view estimated by the attention net.

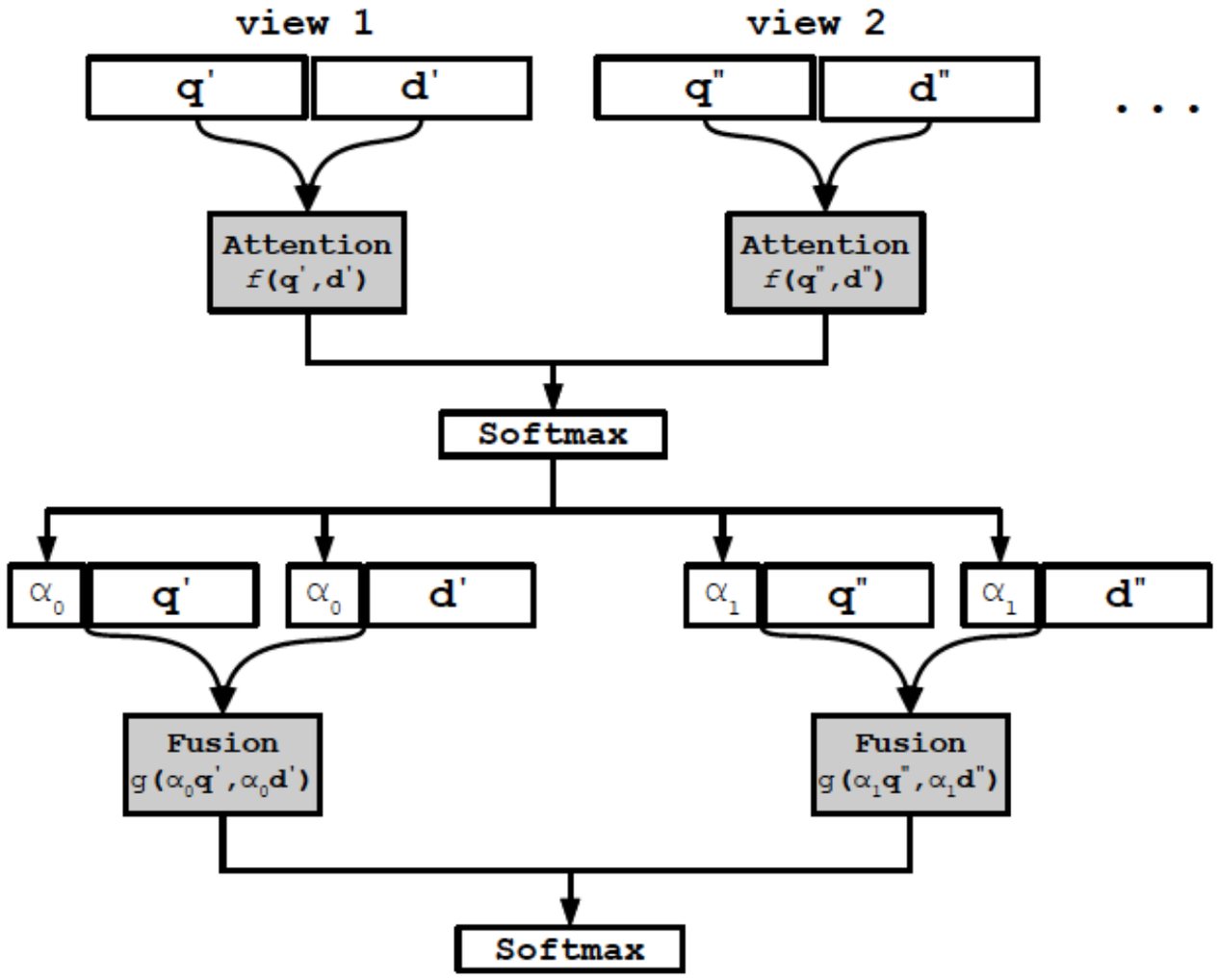
- Views
  1. Text of referral letters and diseases
  2. Medical concepts and codes



Assumption: query-document pair of the more influential view are more similar in the underlying shared space

Ranking Problem

- Attentive Multiview Neural Model (AMNM)
  - $f(.)$ indicates attention network
  - $g(.)$ indicates fusion network
  - a_i indicates the attentive weight of the i^{th} view estimated by the attention net.
- Training/validation
  - For each positive pair in training or validation sets, create a negative pair.
- Testing
  - All the possible patient-disease pairs (9K per patient)



Baselines

- **BM25** (Robertson et al., 1995)
  - An unsupervised approach based on (TF/IDF), and document length.
- **SVMs** (Cortes and Vapnik, 1995)
  - TF/IDF weighted ngrams (n=[1–2])
- **BERT** (Devlin et al., 2019)
  - An attentive deep language model that conditions on left and right contexts.
  - We use BERT models developed for clinical text (Alsentzer et al., 2019).
- **SVM^{rank}** (Joachims, 2002)
  - Empirical risk minimization for ranking problems
  - Scores generated by the above baselines and (IDF-weighted) unigram overlap.
- **PhenoTips** (Girdea et al., 2013)
  - Ontology-based similarity

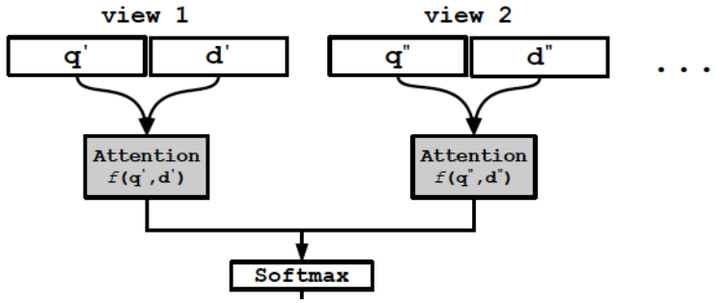
Evaluation

Model	Fusion	MAP	P@5	P@10
SVM^{rank}	text & code	12.9	12.5	12.9
PhenoTips	text & code	15.4	8.3	5.4
$\text{AMNM}_{bert\text{-}bert}$	g^{dot}	**18.9***	14.2	17.5
$\text{AMNM}_{bert\text{-}bert}$	g^{outer}	18.0*	16.7	17.5
$\text{AMNM}_{bert\text{-}bert}$	g^{conv}	16.0*	10.0	12.1
$\text{AMNM}_{bert\text{-}svms}$	g^{dot}	18.4*	**18.3**	**17.9**
$\text{AMNM}_{bert\text{-}svms}$	g^{outer}	17.1*	17.5	17.1
$\text{AMNM}_{bert\text{-}svms}$	g^{conv}	11.4	14.2	13.9

Table 2: Model performance across different fusion functions. The Model column shows the source of representations for text and code views respectively. * indi-

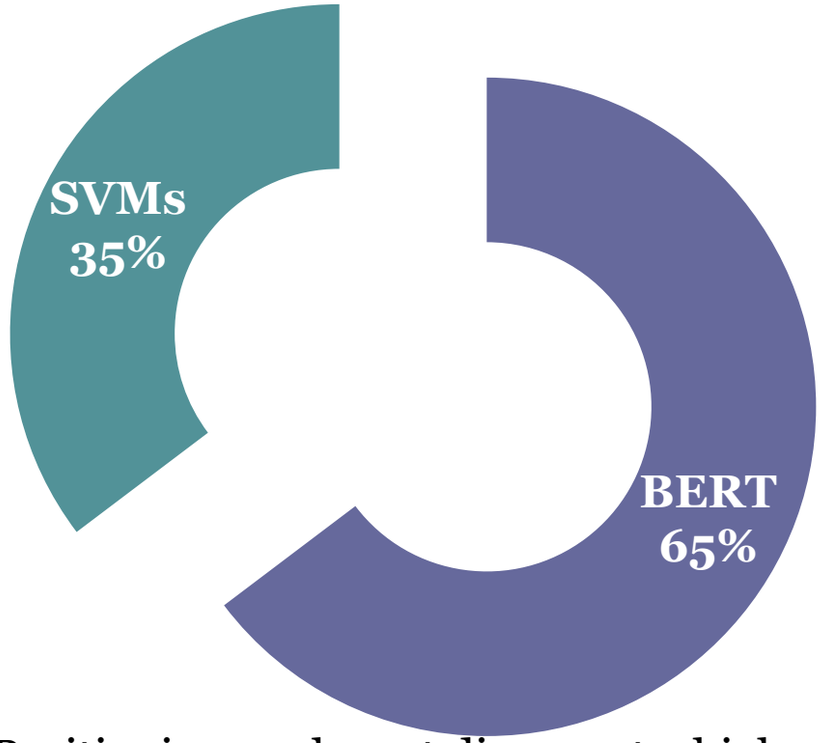
MAP: Mean Average Precision **P@K**: Precision at rank K *: Significant difference

Model Introspection



- Why is our model better?
  - Higher attentive weights assigned to the view that better estimates the relevance score?
    - Separately apply BERT (text) and SVMs (concept) to test data.
    - For *relevant* pairs, evaluate if attention network assigns higher weights to better views.

- 57.7% accurate in prioritizing better views
- Higher weight to ~2K pairs could be the source of improvement.



Positioning a relevant disease at a higher rank

Conclusion & Future Work

- Conclusions
  - Undiagnosed Diseases Network
  - Attentive Multiview Neural Model
    - Attention and fusion networks are important for effective ranking
    - The combination of traditional (SVMs) and recent (BERT) techniques performs best.

- Future Work
  - Adding other views and data modalities
  - Adding (causal) gene-disease and disease-phenotype relations
  - Employing better data sampling strategies

Thank you!

Code: https://clu.cs.uml.edu/tools.html

Acknowledgement

- **NIH Common Fund, U01HG007530**
- **The UDN team members**

HARVARD
MEDICAL SCHOOL

Referral Letters

- 3.9K letters + supporting materials
  - Videos
  - Binary files
  - Password-protected files
  - Not unzippable files
  - Docs, Images, PDFs, Etc.

¯_(ツ)_/¯

Please upload a study recommendation letter from one of the applicant's referring licensed healthcare providers (PDF only). The letter should be on letterhead and signed by the healthcare provider.

Choose File No file chosen

Ignored some applications due to OCR error or file type complexity

Evaluation

	Text View			Code View		
Model	**MAP**	**P@5**	**P@10**	**MAP**	**P@5**	**P@10**
BM25	4.1	5.0	3.8	6.5	8.3	6.3
SVMs	8.8	8.3	8.3	7.7	8.3	8.8
BERT	**15.5**	**12.5**	11.7	**10.8**	**13.3**	**10.8**
SVM^{rank}	12.1	9.2	**12.5**	8.5	8.3	8.6

Table 1: MAP, P@5 and P@10 performance of baselines (in percentages) on text and code views.