DATA SCIENCE IN BIOMEDICINE

Prof. Dr. Holger Fröhlich
Head of AI & Data Science Group, Deputy Head of Department of Bioinformatics
Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)

Campus Birlinghoven

- One of Germany‘s largest research centers for applied computer science and mathematics

- 800 employees
  - 600 scientists
  - 200 students and trainees

- 3 research institutes
  - Scientific Computing and Algorithms (SCAI)
  - Applied Information Technology (FIT)
  - Intelligent Analysis and Information Systems (IAIS)

AI & Data Science Group @Fraunhofer SCAI
Mission: Bringing Better Treatments to the Right Patients

Methodologies, e.g.

- Omics
- Longitudinal clinical
- Real World Data
  - EHRs
  - Device data
- Biol. sequences & text

>150 publications

Hybrid AI: combining human knowledge with AI

(Generative) time series models

Multi-modal, multi-scale data fusion

AI & Data Science

Application focus

Precision medicine

Systems Medicine & Drug Discovery

- Pharma
- Biotech
- Insurances
- Public Health

Our value chain:

Biological system reconstruction & modeling → Drug target identification → The right drug for the right patients → Clinical trial design: simulation of synthetic cohorts

https://www.scai.fraunhofer.de/de/geschaeftsfelder/bioinformatik/forschungsfelder/AI-DAS.html

Projects and Partners

Systems Medicine

- ADIS (JNPD, coordinator)
- Industry collaboration

Precision Medicine

- DIGIPD (ERA PerMed, coordinator)
- The Virtual Brain Cloud (Horizon)
- RADAR-AD (IMI)
- PsychSTRATA (Horizon)
- Industry collaboration

Clinical Trial Design, Synthetic Cohorts

- NFDI4Health (DFG)
- Industry collaboration

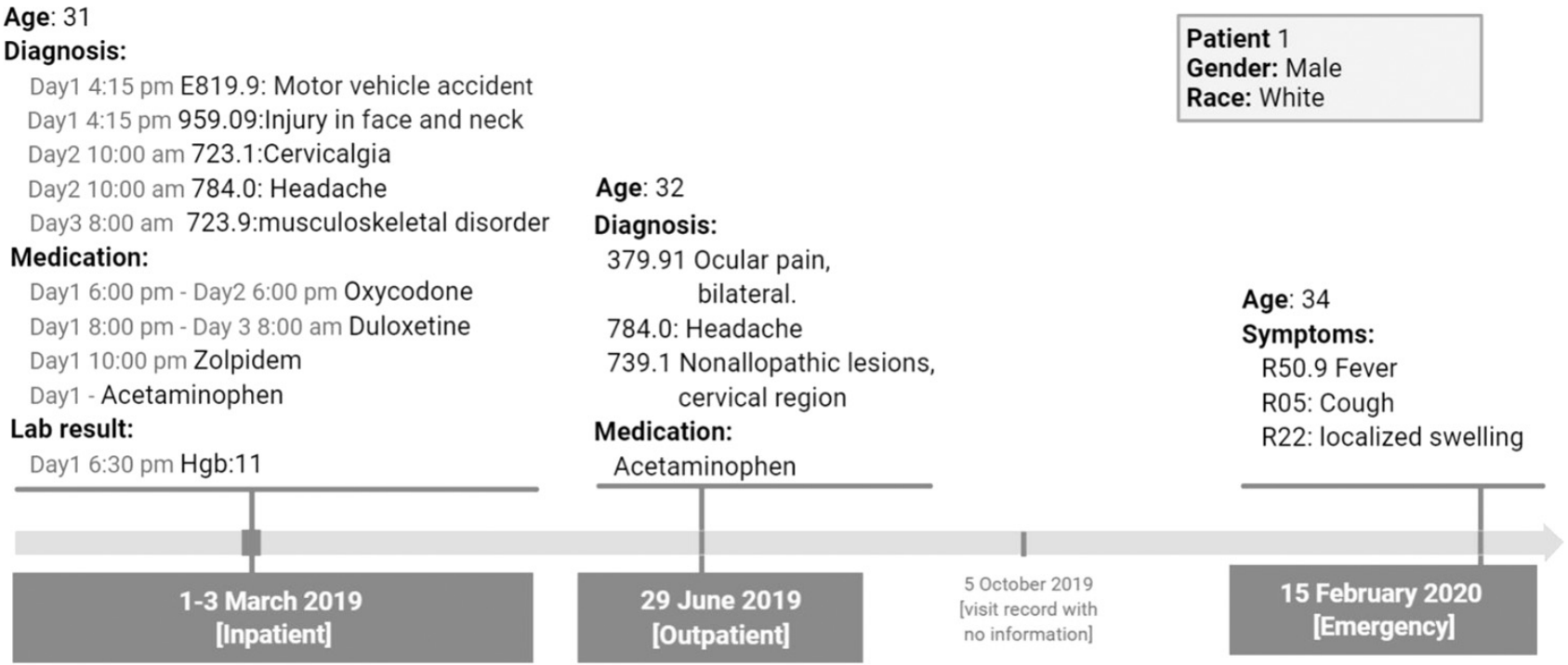
Decision Support for Public Health

- AIOLOS (BMWK)
- Real4Reg (Horizon)

Example 1: Real-World Data (RWD) in Medicine

- RWD reflect the situation of patients in medical practice
  - Contrast: medical research data (e.g. clinical studies)
- Different types of RWD, e.g.
  - **Electronic health records (EHRs)**
  - Laboratory and diagnostic data (e.g. EEG, ECG, CT scans) in routine care
  - Social media and internet searches
  - Data collected via smartphones and other digital health applications
- NONE of these data have been collected for AI/ML purposes

Example of structured EHR Data



The Challenge: AI for Real-World Data

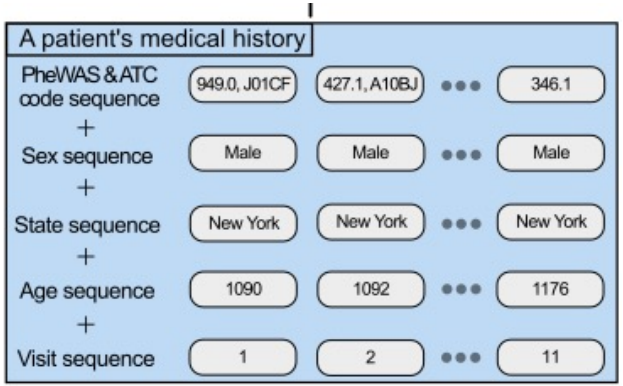
How can we effectively leverage large scale structured EHR data to build AI/ML models?

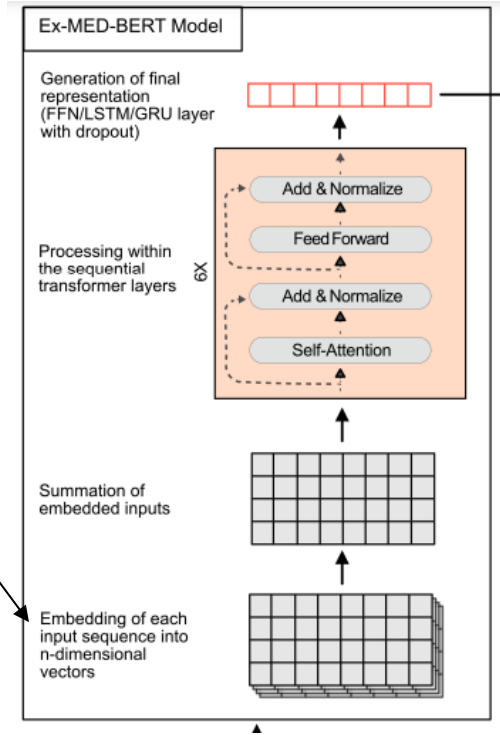
- Irregular, static + longitudinal time series data
- Mostly discrete codes, but can also contain quantitative values

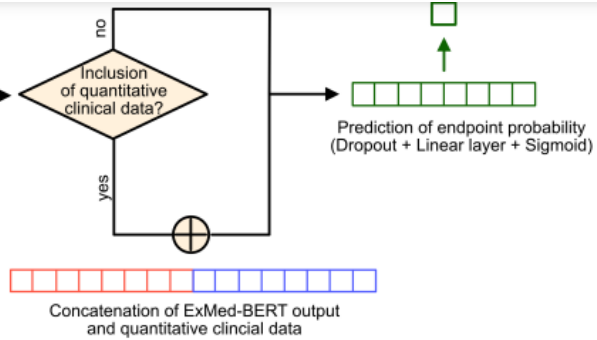
Different possible use cases

Type of model	Purpose	Value
Risk model (classifier or time-to-event)	Identify patients with elevated disease risk	Earlier diagnosis, disease prevention measures
Representation learning, clustering	Identify patients with similar health trajectories	Subgroup specific, targeted therapies
Classifier	Real-world drug effectiveness and safety	Market understanding, treatment recommendation

ExMed-BERT: A Transformer Model for EHRs







Pre-training with ~1 Billion records from 3.8 Million patients

Vector representation (d = 192)

Linden, …, Fröhlich, Frontiers in Artificial Intelligence, 2021

Lentzen, …, Fröhlich, https://www.medrxiv.org/content/10.1101/2022.11.29.22282632v1

Example Application: Predicting COVID-19 Disease Severity

- Endpoint: acute respiratory manifestation within 3 weeks after COVID-19 infection
  - Adjustment for confounding effects of age and sex via inverse probability of treatment weighting (IPTW)
  - 1y medical history for each patient
- ExMed-BERT shows better performance than competing methods (AUC ~80)
  - Addition of quantitative clinical measures helps

	IBM Explorys Therapeutic Dataset
Entire dataset (unfiltered)	4,563,769
Pre-training cohort	3,478,438
Fine-tuning cohort	80,211
ARM patients	10,743
Patients with quantitative clinical data	23,949

Making Models Explainable

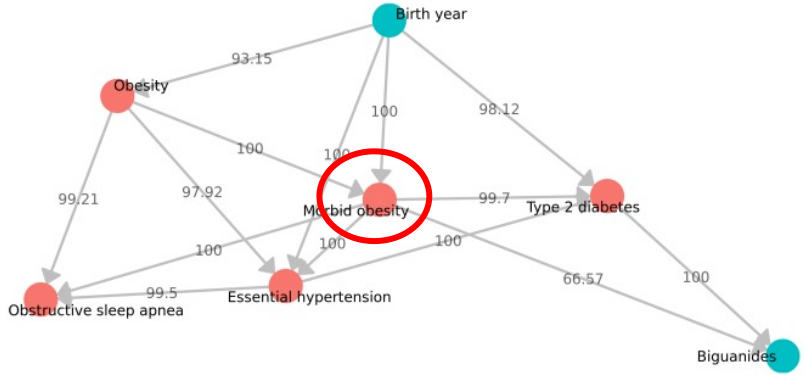
Fine tuned ExMed-BERT model

Integrated Gradients

Bayesian Network of top features

- Bootstrapped structure learning

Statistical testing of marginal effects for causal understanding



Feature Names	Corrected p-value
Type 2 diabetes	<0.0001
Heart failure with reduced EF	<0.0001
Shortness of breath	<0.0001
Constipation	<0.0001
Morbid obesity	<0.0001
Screening for malignant neoplasms	<0.0001
Dementias	<0.0001
Chronic airway obstruction	0,0003
Cough	0,0018
Screening for infectious and parasitic diseases	0,0083
Congestive heart failure (CHF) NOS	0,0119

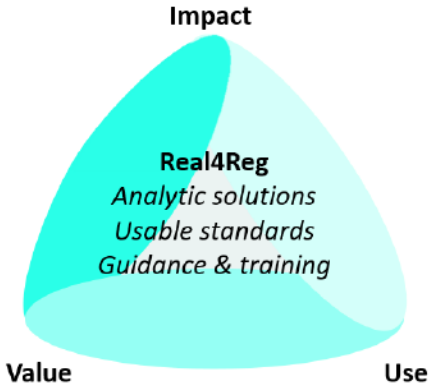
Summary: AI for Real-World Data

- Custom-made large language model (transformer architecture) for dealing with structured EHR data
- Pre-trained transformed model for structured EHRs
  - Uses medications, diagnoses and demographics
  - Combination with quantitative data possible
  - Will be made accessible to community: https://zenodo.org/record/7324178#.Y-n7P3bMK3A
- Innovative strategy to make models explainable

AI for Regulatory Decision Support: Real4Reg (2023 – 2027)

3 objectives:

- Enable use of real-world data
- Establish the value
- Impact RWD use and analyses through training



Data Science related objectives

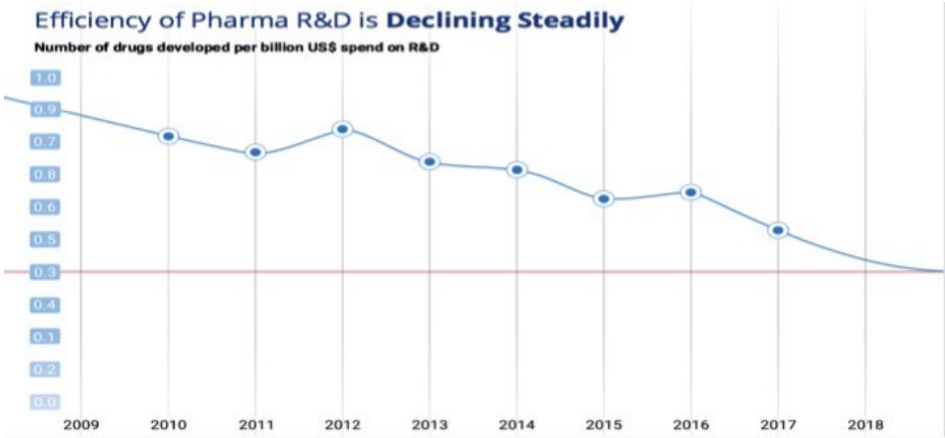
- Harmonize meta-data provided by different partners to OMOP CDM
- Develop workflow to subset and display RWD
- Workflow for synthetic control arms
- AI/ML algorithms for predicting drug effectiveness and safety
- AI algorithms for synthetic data generation

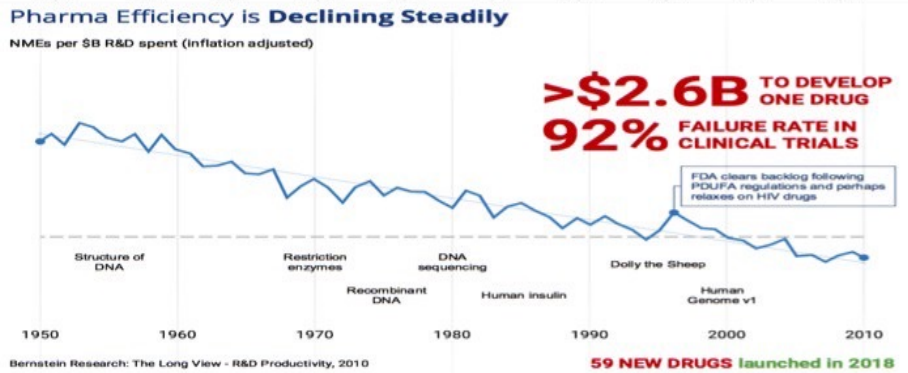
Bundesinstitut für Arzneimittel und Medizinprodukte

https://www.aerzteblatt.de/nachrichten/140328/EU-Projekt-Kuenstliche-Intelligenz-soll-Arzneimittelregulierung-verbessern

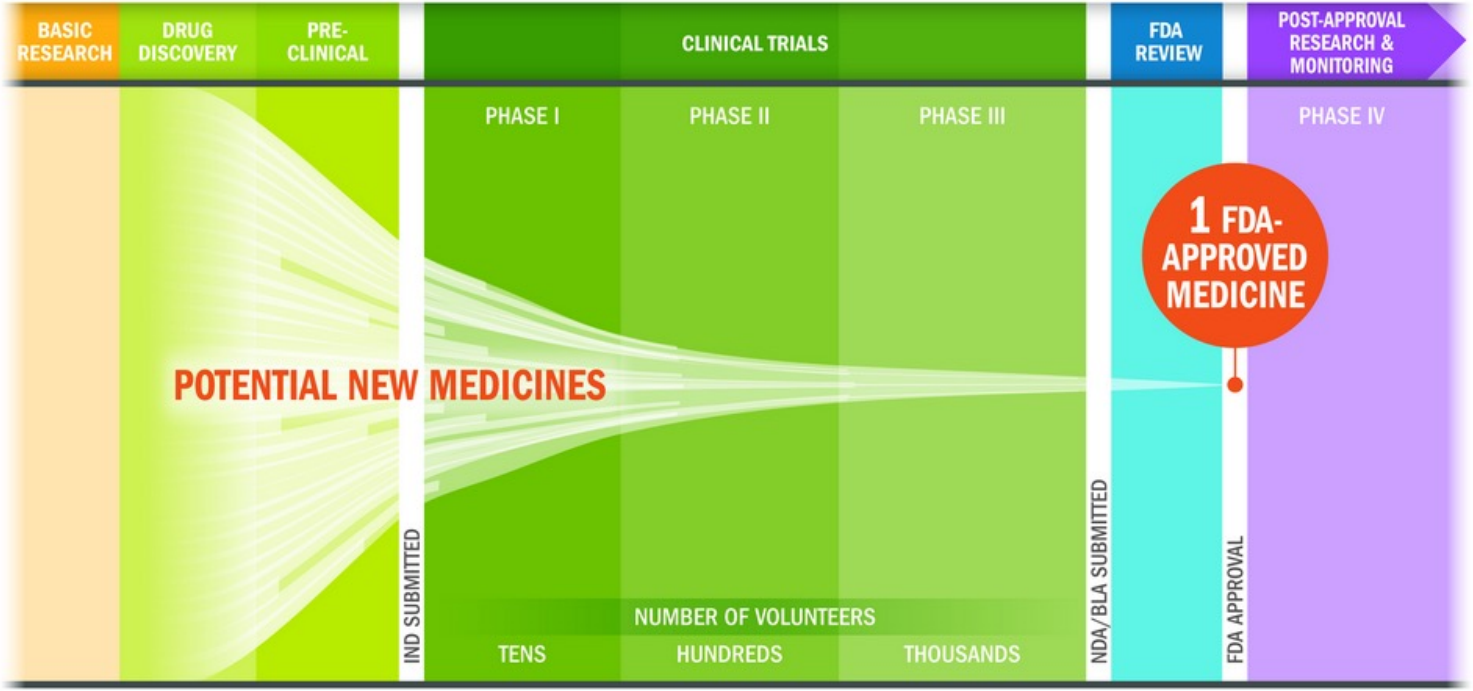
Example 2: AI in Drug Development

- Challenge: > 90% of clinical studies fail
- Top reasons:
  - Lack of efficacy
  - Unwanted side effects
- Could AI help?





The Drug Development Process

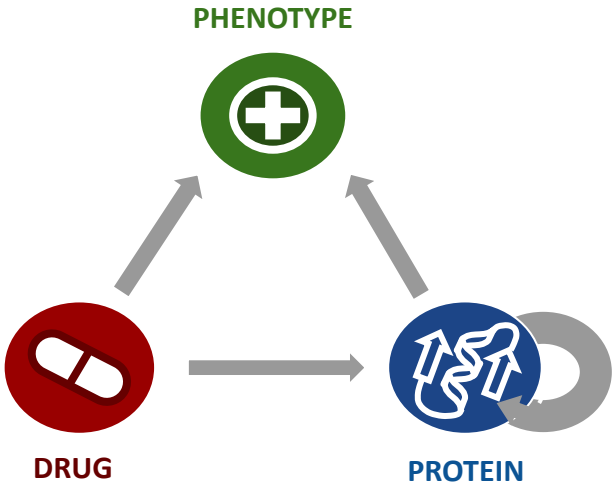


~$2.6B

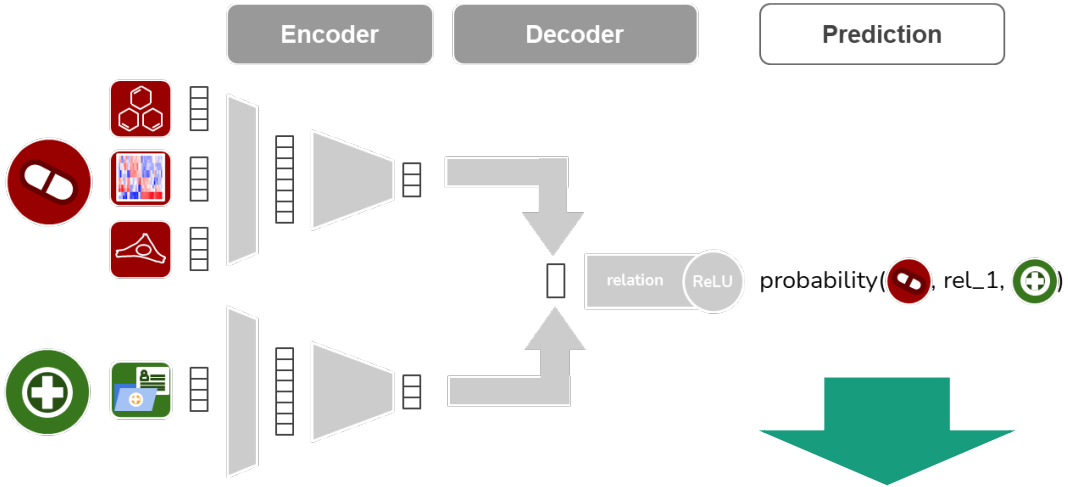
Source: PhRMA adaptation based on Tufts Center for the Study of Drug Development (CSDD) Briefing: "Cost of Developing a New Drug," Nov. 2014. Tufts CSDD & School of Medicine., and US FDA Infographic, "Drug Approval Process," http://www.fda.gov/downloads/Drugs/ResourcesForYou/Consumers/UCM284393.pdf (accessed Jan. 20, 2015).

AI for Prediction of Unwanted Side Effects of Drugs

Knowledge + **Multimodal Data** + **AI (Graph Neural Network)**



- ~30,000 entities
- ~400,000 relations

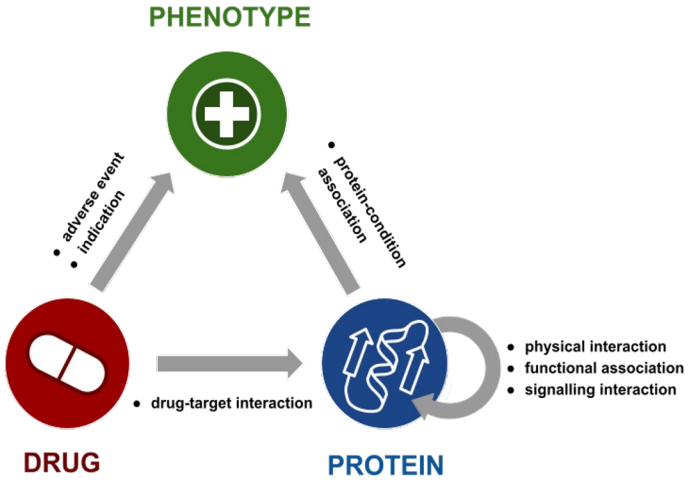


- compound → side effects
- target → side effect

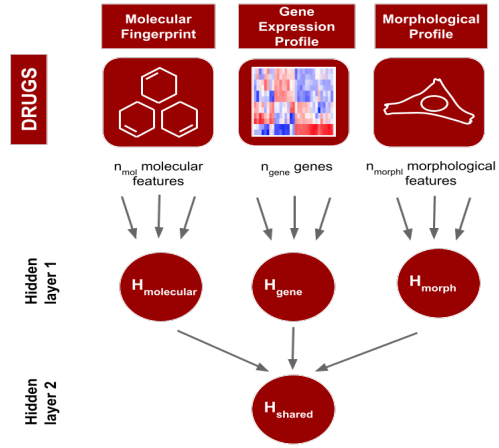
Sharma, …, Fröhlich, submitted
Krix,…, Fröhlich, submitted

Some Details

reactome PheWAS DRUGBANK ClinicalTrials.gov IntAct Open Targets DisGeNET SIDER IUPHAR KEGG STRING BioGRID

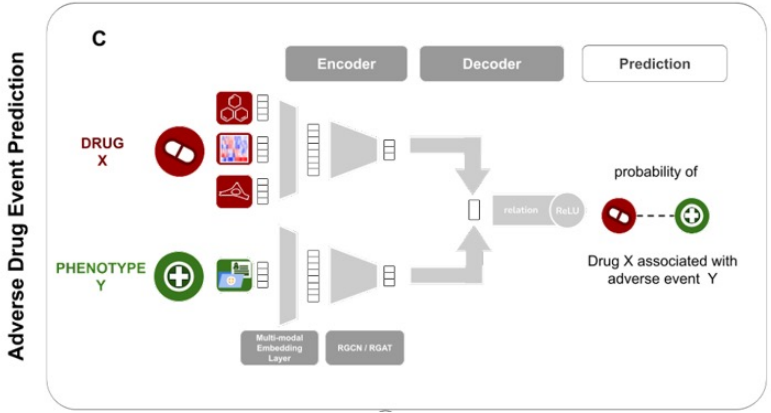


MultiGML Graph Neural Network



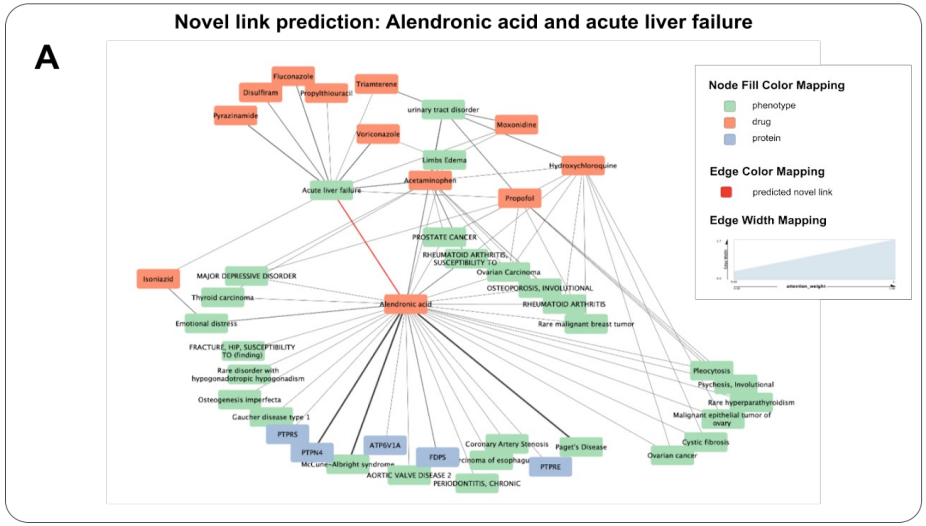
Multi-modal embedding layer

Relational Graph Attention Network



Example Results

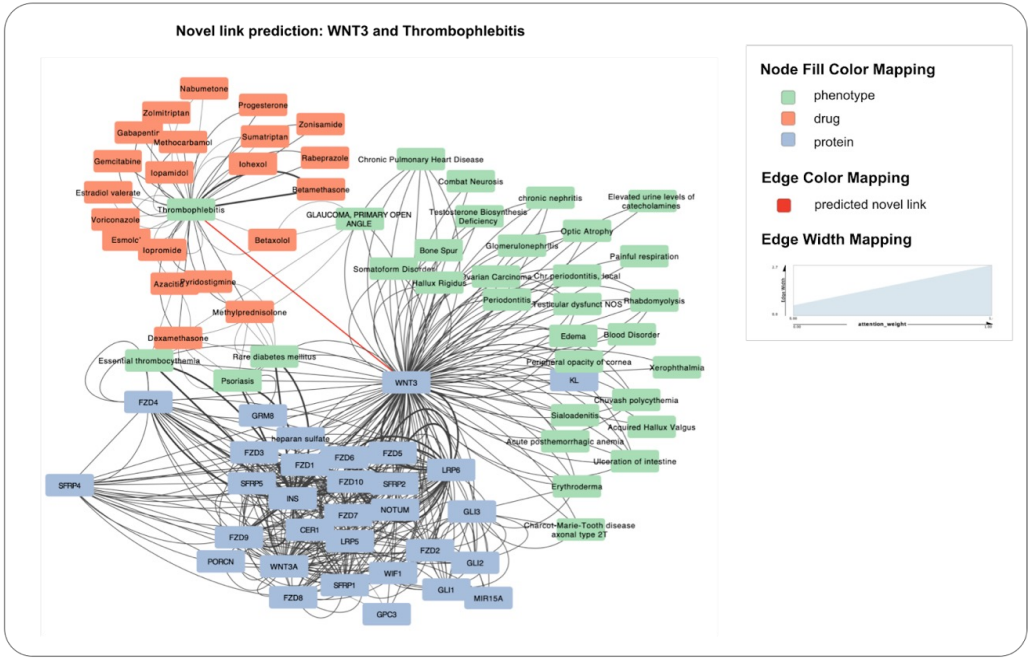
- Superior prediction performance of MultiGML compared to traditional knowledge graph embedding methods
  - AUC ~1 for side effect prediction; AUC ~90% for general phenotype prediction
- Integrated gradient approach to make MultiGML predictions explainable
- Novel prediction supported by literature



Fraunhofer

MultiGML as a Model for Target De-risking

- Thrombophlebitis is an inflammation of a vein associated with a blood clot
- MultiGML predicts association of specific protein target (WNT3) with this phenotype
- Strong support by biological literature



Fraunhofer

Summary: AI for Prediction of Unwanted Side Effects and Target Prioritization

MultiGML is a novel Graph Neural Network architecture integrating domain knowledge and multimodal data

MultiGML demonstrates high prediction performance for predicting novel relationships between compounds and phenotypes as well as targets and phenotypes

MultiGML could be used during target selection and compound development phases

- Ongoing project together with pharma partner

Conclusion

AI plays an increasing role in healthcare

- Multitude of questions and data types
- Vibrant research area

Examples:

- AI for Real-World Data Analysis
- AI in early Drug Development

AI & Data Science Team @Fraunhofer SCAI

Analysis of health related data

- Multi-omics
- Longitudinal clinical data
- Real World Data
- Biological sequences & text

Developing problem specific algorithms & solutions

- Hybrid AI: combining human knowledge with AI
- (generative) time series modeling
- multi-scale, multi-modal data fusion

Working within international networks of excellence

RADAR-AD
Real4Reg
adis
AIOLOS
Artificial Intelligence ToOLs for Outbreak Detection and response
DIGIPD
PsychSTRATA
nfdi 4 health

holger.froehlich@scai.fraunhofer.de

Fraunhofer

Quantitative Clinical Data

Only features with <60% missingness included

Require features to be recorded 2 weeks prior diagnosis

8 features are left

- Weight
- BMI
- Body surface areas
- Body height
- Temperature
- Heart rate
- Diastolic and systolic blood pressure

Prediction Performances Ex-MedBERT

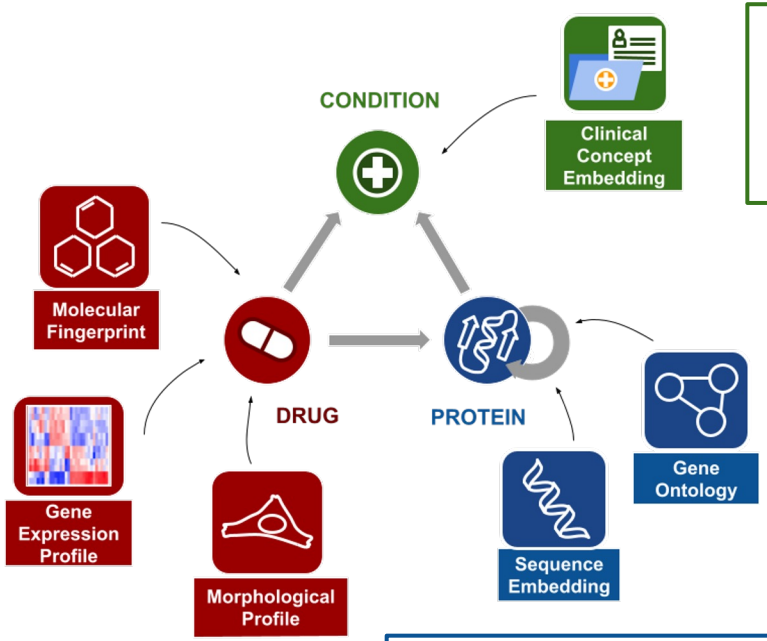
Model	AUROC [%]	AUPR [%]
RF	73.4 [72.6, 74.3]	29.1 [27.6, 30.7]
+ Quant	77.7 [76.9, 78.6]	34.7 [33.0, 36.4]
subset w/o missingness	68.6 [67.1, 70.1]	40.1 [37.8, 42.6]
subset w/o missingness + Quant	70.2 [68.8, 71.6]	42.1 [39.7, 44.5]
XGB	72.4 [71.5, 73.3]	28.2 [26.7, 29.7]
+ Quant	77.7 [76.9, 78.5]	35.5 [33.8, 37.3]
subset w/o missingness	67.8 [66.3, 69.2]	38.8 [36.5, 41.2]
subset w/o missingness + Quant	70.6 [69.2, 72.0]	42.2 [39.7, 44.7]
ExMed-BERT-FFN	77.5 [76.7, 78.4]	38.2 [36.4, 40.0]
+ Quant	77.7 [76.8, 78.5]	38.1 [36.3, 39.8]
subset w/o missingness	67.6 [66.1, 69.1]	39.3 [36.9, 41.6]
subset w/o missingness + Quant	70.1 [68.7, 71.6]	41.9 [39.5, 44.4]
ExMed-BERT-GRU	77.7 [76.8, 78.6]	36.7 [35.0, 38.4]
+ Quant	**79.8 [78.9, 80.6]**	38.7 [36.9, 40.4]
subset w/o missingness	70.7 [69.3, 72.1]	42.8 [40.2, 45.4]
subset w/o missingness + Quant	72.0 [70.5, 73.5]	44.7 [42.1, 47.3]
ExMed-BERT-LSTM	77.7 [76.9, 78.6]	37.6 [35.9, 39.4]
+ Quant	78.4 [77.4, 79.3]	39.3 [37.4, 41.0]
subset w/o missingness	71.8 [70.5, 73.3]	**45.0 [42.4, 47.4]**
subset w/o missingness + Quant	71.4 [70.0, 72.8]	43.3 [40.8, 45.8]

Feature Generation and Annotation

Morgan count fingerprint (RDKit) with radius = 2 (Landrum et al., 2010)

Consensus transcriptional signatures (Himmelstein et al., 2016) from the L1000 dataset (Duan et al., 2014) since each compound has ben assayed across multiple cell lines and dosages

Drug perturbation effects in a cell culture experiment on cell morphology changes (Schreiber, 2014)



Medical concept embeddings *cui2vec* (Beam et al., 2020) were generated on the basis of clinical notes, insurance claims and biomedical full text articles

Binary Gene Ontology (GO) Fingerprint for biological processes using the Gene Ontology Resource (Ashburner et al., 2000)

Structural information on protein sequences from the deep learning model ESM-1b Transformer (Rives et al., 2021)

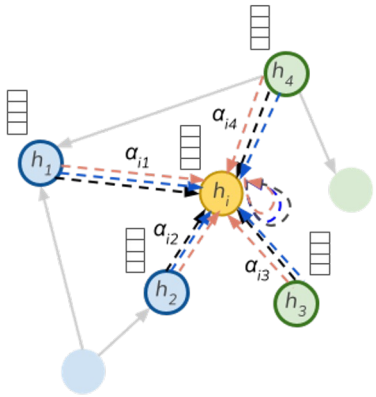
MultiGML-RGAT
Relational Graph Attention Network (RGAT)

- Self-attention on nodes:
  - Attention coefficients $E^{(r)}_{i,j}$ for each relation type are computed with shared attention mechanism
  - Attention coefficients are normalized across all choices of *j* via the softmax function
- Attention mechanism *a* is a single-layer feedforward neural network, parametrized by a weight matrix *W,* feeding into a Leaky ReLU
- Propagation model for a single node update in a multi-relational graph:

$$h_i^{(l+1)} = \sigma\left(\sum_{r \epsilon R}\sum_{j \epsilon N_i^r} \alpha_{i,j}^r h_i (W_r)^T\right)$$

$$E_{i,j}^{(r)} = a(W^{(r)} h_i, W^{(r)} h_j)$$

$$\alpha_{i,j}^{(r)} = softmax_j\left(E_{i,j}^{(r)}\right)$$

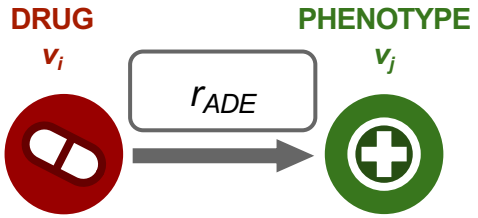


Neighborhood aggregation with attention mechanism.

(Veličković et al., 2017)
(Busbridge et al., 2019)

Bilinear Decoder

- Decoder uses entity embeddings to decode relations in the knowledge graph
- Calculate score for entities v_i and v_j, connected by relation r
- Use bilinear form on entity embeddings h_i and h_j with trainable relation-type specific matrix M_r
- Apply sigmoid function for a score between [0 , 1]



$$score_{i,j}^{(r)} = score(v_i, r, v_j) = \sigma(h_i^T M_r h_j)$$

Model Training Strategy

- Training of both MultiGML variants (-RGCN and -RGAT) on the knowledge graph with input features
- Stratified data split into approx. 70% training, 10% validation set and 20% test on relation types

Data Set	Training	Validation	Testing
Number of Samples	604.903	67.212	168.029

- All real existing relations provide positive examples
- Negative sampling with k = 1
  - Randomly corrupting the target entity for each source entity according to uniform distribution
- Hyperparameter Optimization with optuna using the Tree Parzen Estimator (Bergstra et al., 2011; Akiba et al., 2019)
  - 50 trails for each MultiGML variant
- Training for 100 epochs with best hyperparameters

Prediction Performances MultiGML: Adverse Event Prediction

Model	AUROC	AUPR	Precision@30
TransE	0.293	0.389	0.233
RotatE	0.943	0.915	0.967
ComplEx	0.884	0.934	**1.0**
DistMult	0.963	0.966	**1.0**
DeepWalk	0.575	0.604	0.567
Random Forest	0.512	0.164	0.464
MultiGML-RGCN (basic)	**1.0**	**1.0**	**1.0**
MultiGML-RGAT (basic)	**1.0**	**1.0**	**1.0**
MultiGML-RGCN (multimodal)	**1.0**	**1.0**	**1.0**
MultiGML-RGAT (multimodal)	0.980	0.982	**1.0**

Prediction Performances MultiGML: Phenotype Prediction

Model	AUROC	AUPR	Precision@100
TransE	0.735	0.674	0.533
RotatE	0.723	0.680	0.710
ComplEx	0.843	0.770	0.790
DistMult	0.848	0.767	0.776
DeepWalk	0.654	0.630	0.761
MultiGML-RGCN (basic)	**0.898**	**0.832**	**0.850**
MultiGML-RGAT (basic)	0.897	0.831	0.841
MultiGML-RGCN (multimodal)	0.897	**0.832**	0.846
MultiGML-RGAT (multimodal)	0.892	0.826	0.827

Making Models Explainable
Integrated Gradients

Captum

- Important to explain which features of an entity influence the model's predictions from an application perspective
- Used the **Integrated Gradients** method (Sundararajan et al., 2017) implemented by captum.ai
  - Axiomatic attribution method
  - Represents the integral of gradients w.r.t. inputs along the path from a given baseline to input
- Integrated gradient along the i^{th} dimension from baseline x_0 to input x :

$$\text{IntegratedGrads}_i(x)^{\text{approx}} := (x_i - x_i') \times \int_{k=1}^{m} \frac{\partial F(x' + \frac{k}{m} \times (x - x'))}{\partial x_i} \partial\alpha$$

with scaling coefficient α, number of steps *m* and function *F* representing the MultiGML model