The role of semantics and ontologies in interpretable machine learning

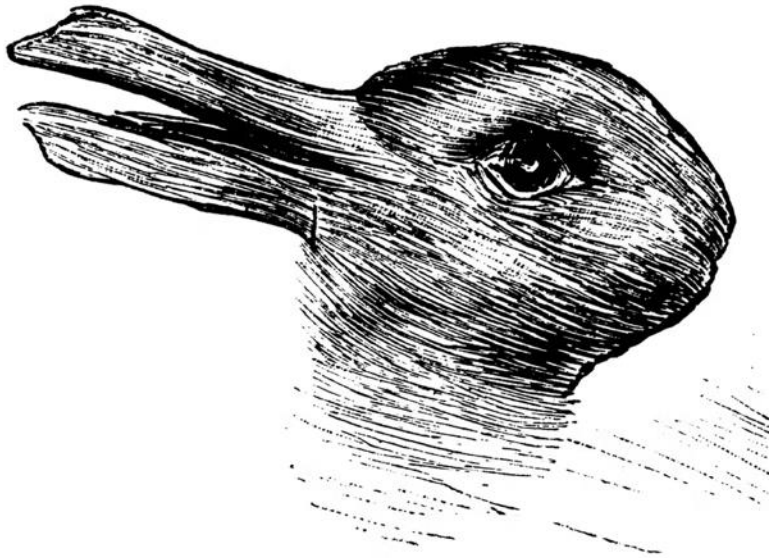
Janna Hastings,

OvGU Magdeburg, UCL, EPFL

Assistant Professor of Medical Knowledge and Decision Support,
Medical Faculty, University of Zurich & School of Medicine, University of St. Gallen
(from August 2022)

Journal Club Genomics and Digital Health, University of Geneva
Campus Biotech, 25 May 2022

There are ‘bottom-up’ and ‘top-down’ processes in human mental functioning



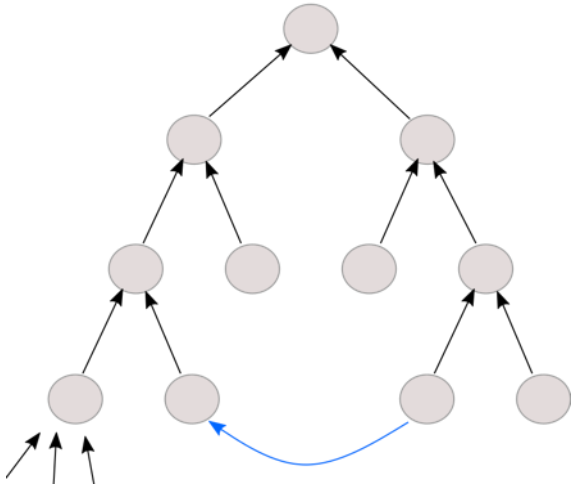
Jastrow, J. Popular Sci. Monthly **54**, 299-312, 1899.

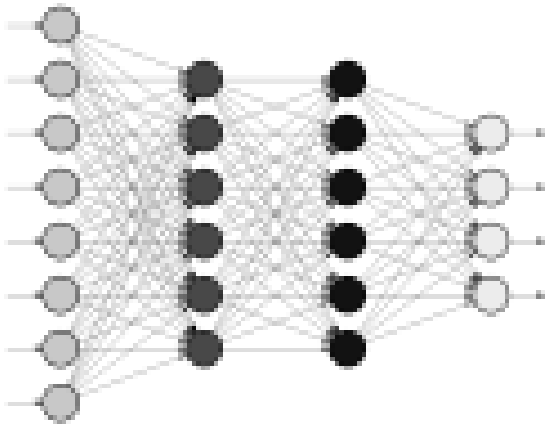
Bruner, J. S., & Minturn, A. L. (1955).
Journal of General Psychology, 53, 21–28.

Bugelski, B. R., & Alampay, D. A. (1961).
Canadian Journal of Psychology, 15(4), 205–211.

that are tightly **integrated**

There are also (analogues to) ‘bottom-up’ and ‘top-down’ processes in machine learning and reasoning





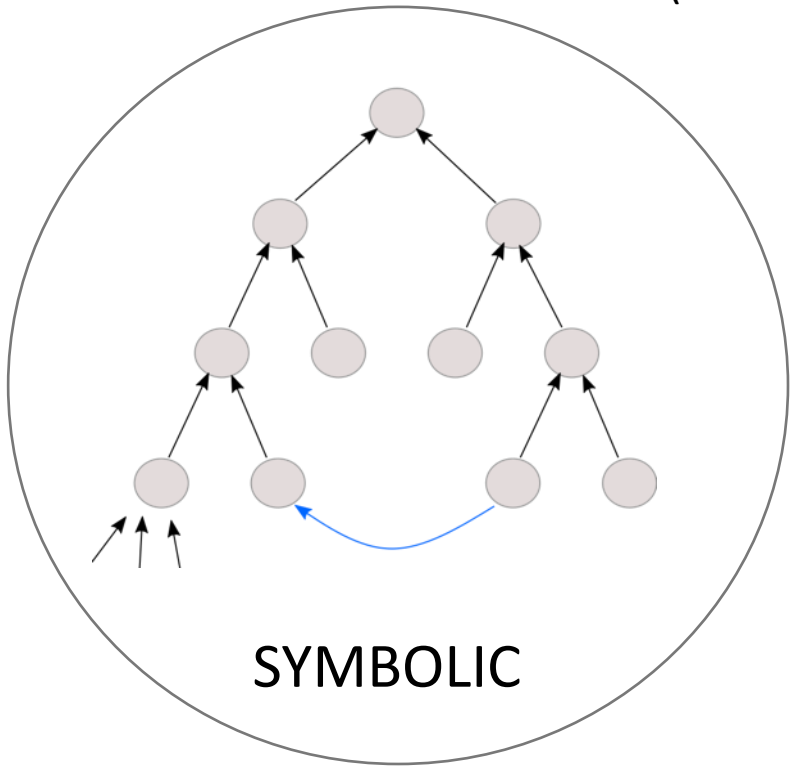
Ontologies, Semantics, Logical Inference, Automated Reasoning, …

Deep Neural Networks, Bayesian Learning, Signal processing, …

that are (mostly) **not** integrated

Two different (computational) ‘worlds’

“Good Old Fashioned AI” (1970s)



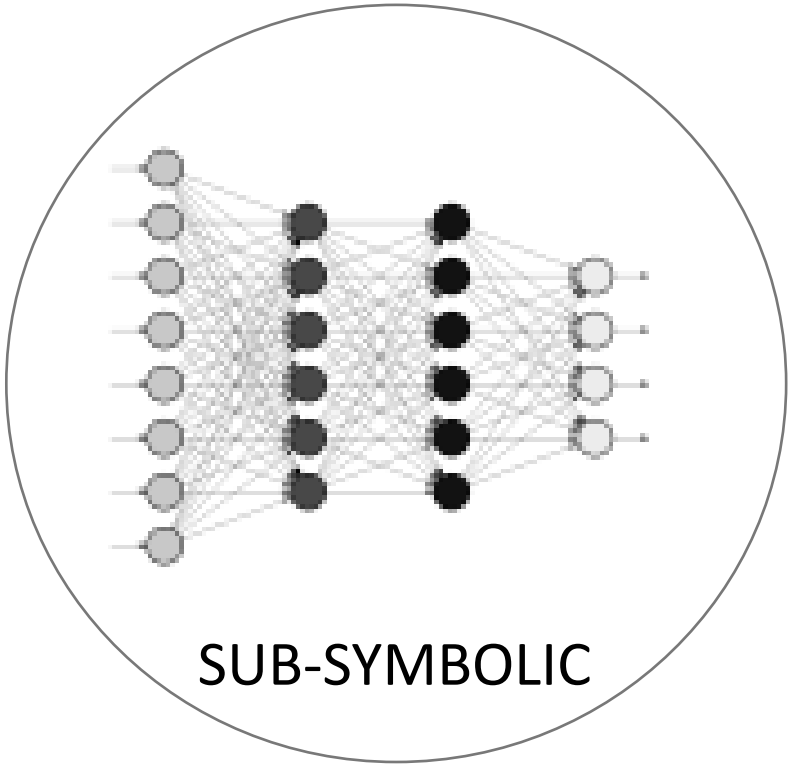
SYMBOLIC

Rigid, Logical, Rules-driven

+ Provable outcomes, Transparent

- Slow, Hard to maintain, Vulnerable to inconsistencies

“The Singularity is Near” (2010s)



SUB-SYMBOLIC

Quantitative, Associative, Pattern-driven

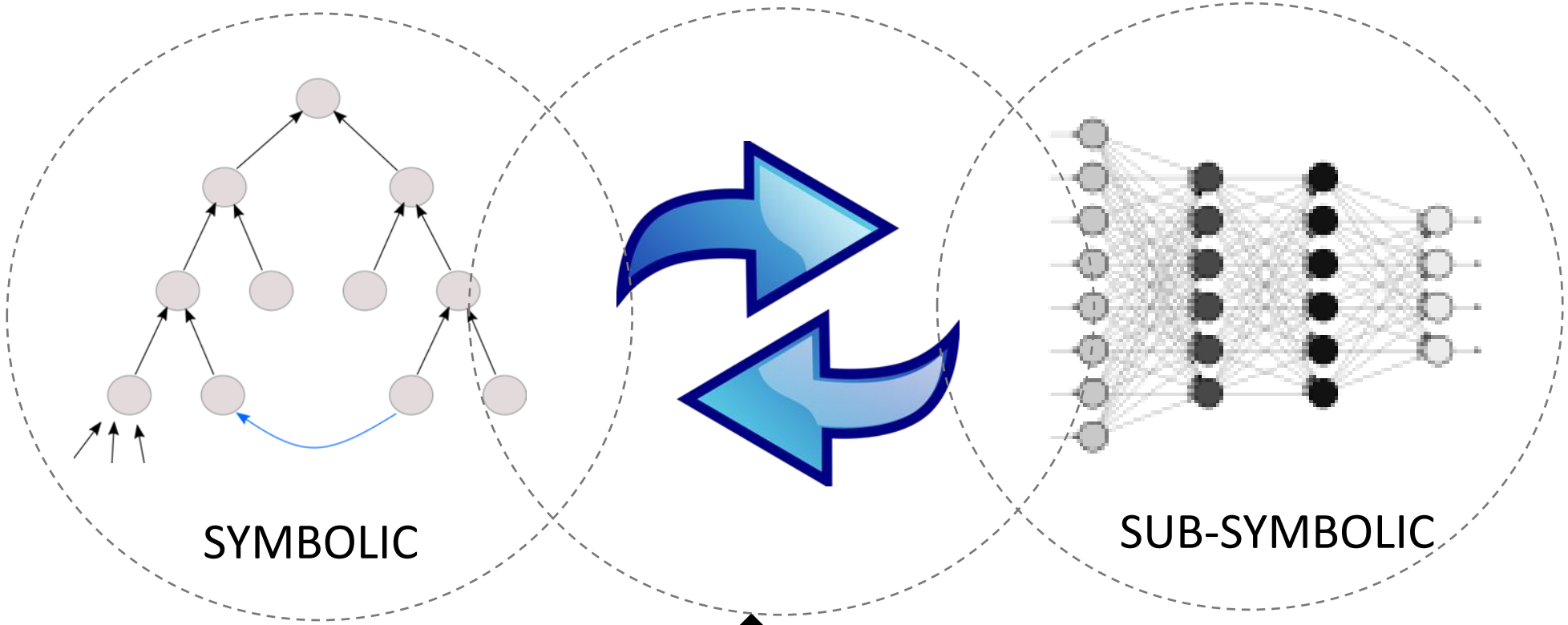
+ Scalable, Robust to inconsistencies

- Vulnerable to bias, errors and attacks, Impenetrable

Semantics and interpretability are **essential** for many applications of ML in medicine

- Interpretability for decision-making responsibility / accountability
- Logic for verifiability (approximately correct not always good enough, may need provable correctness)
- Small(ish) data, need performance even when don't have massive datasets (lots of reasons – availability, privacy, decentralization and climate)

How do we build systems with the best of both?

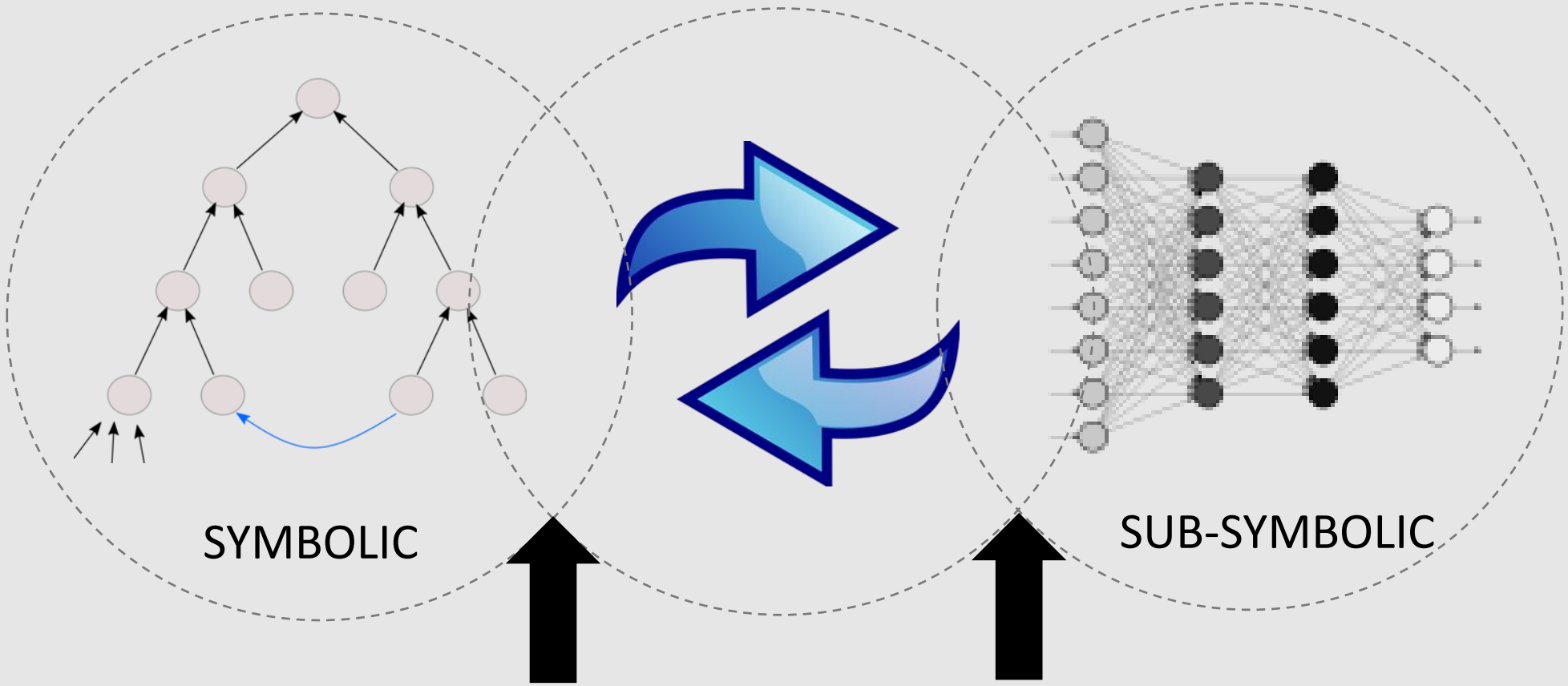


Rigid, True or False, Based on Logic
+ Provable outcomes, Transparent

Quantitative, Associative, Based on Statistics
+ Scalable, Robust to inconsistencies

Data-centric AI / Neuro-symbolic AI / Semantic AI / 'Informed' ML / 'Broad' AI

Two case studies: (1) evidence synthesis and (2) (automated) semantic annotation of chemicals



Interpretable prediction of the outcomes of smoking cessation interventions

Interpretable semantic annotation of chemical structures to a chemical ontology

UNIVERSITY OF ABERDEEN

UNIVERSITY OF CAMBRIDGE

IBM Research

UCL

Centre for Behaviour Change

Harnessing cross-disciplinary expertise to address social, health and environmental challenges.

Interpretable Prediction of the Outcomes of Smoking Cessation Interventions

The Human Behaviour-Change Project

A Collaborative Award funded by the wellcome trust

HB CP

Human Behaviour-Change Project

COMMENT | 15 December 2021

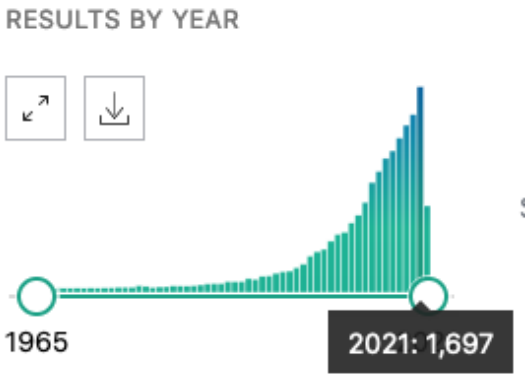
Decision makers need constantly updated evidence synthesis

Nature 600, 383-385 (2021)

Fund and use 'living' reviews of the latest data to steer research, practice and policy.

Julian Elliott , Rebecca Lawrence , Jan C. Minx , Olufemi T. Oladapo , Philippe Ravaud , Britta Tendal Jeppesen , James Thomas , Tari Turner , Per Olav Vandvik & Jeremy M. Grimshaw

The Evidence 'Wave'



'behaviour change intervention'

~ 5 / day, every day
and growing exponentially

The raw data: RCT reports in PDF format

Table 1. Baseline characteristics of the trial participants.

	Total ($N = 572$)	CASC ($N = 285$)	Control ($N = 287$)
Age	46 ± 11	47 ± 11	45 ± 12
Age ranges			
<40 years	171 (30)	75 (26)	96 (33)
40–49 years	179 (31)	97 (34)	82 (29)
50–59 years	159 (28)	82 (29)	77 (27)
≥60 years	63 (11)	31 (11)	32 (11)
Male sex	426 (75)	216 (76)	210 (73)
Smoking history			
Years of smoking	25 (18–33)	25 (19–32)	24 (17–33)
Cigarettes per day	20 (15–20)	20 (15–20)	20 (15–20)
Pack-years	20 (14–30)	21 (14–30)	20 (13–31)
Exhaled CO (ppm)	17 ± 11	17 ± 10	18 ± 11
TDS score	7.7 ± 1.5	7.7 ± 1.4	7.8 ± 1.5
FTND	5.3 ± 2.1	5.2 ± 2.0	5.3 ± 2.1
Comorbidities			
Cardiovascular diseases	93 (16)	45 (16)	48 (17)
Respiratory diseases	93 (16)	48 (17)	45 (16)
Psychiatric diseases	31 (5)	12 (4)	19 (7)
Prescribed Medication			
Varenicline	454 (79)	227 (80)	227 (79)
Nicotine patch	114 (20)	56 (20)	58 (20)
No medication	4 (1)	2 (1)	2 (1)

Data include mean ± standard deviation, number (%), or median (interquartile range) scores.
CO carbon monoxide, *FTND* Fagerström test for nicotine dependence, and *TDS* tobacco dependence screener.

novel smartphone system for smoking cessation integrated with a mobile exhaled CO checker to supplement standard treatment with face-to-face behavioral support and pharmacotherapy; (2) it

In conclusion, a novel digital thera
significantly improved a long-term CAR
conjunction with standard smoking
patients with nicotine dependence. Digi
cessation may be a promising strateg
prevalence worldwide, and future resear
warranted.

METHODS

Study design

This was a multi-center, randomized, controlle
detailed study protocol was presented elsew
were recruited from 31 smoking cessation clin
2017 to January 2018, and allocated 1:1 to th
and the control group. The intervention group
the control group used the control app, each fo
12-week standard smoking cessation treatm
varenicline maintenance therapy for 24 week
whether or not the effectiveness of the CASC s
after discontinuation of using the app, and we
24 weeks and evaluated CARs up to 52 weeks.
compliance with CONSORT statements. The pro
consent forms were reviewed and approved
Board at Keio University School of Medicine a
This trial was registered at the University Ho
Network (UMIN) Clinical Trials Registry (UMIN0

Participants

We recruited nicotine-dependent adults who
receive smoking cessation treatment under th
Insurance program. Physicians at each clinic screened the patients and obtained written informed consent. The baseline data were collected using a self-administered questionnaire at the initial visit. The program consisted of five visits during a 12-week period, and counseling and pharmacotherapy, including nicotine patch or varenicline were provided by physicians[6]. Inclusion criteria were as follows: adult current smokers who (1) were diagnosed with nicotine dependence (tobacco Dependence Screener [TDS] score ≥5 points)[29]; (2) had a smoking history of pack-years ≥10; (3) intended to quit smoking immediately; (4) agreed to participate in a

Table 2. Changes in MPSS, FTCQ-12, and KTSND scores from baseline to weeks 24 and 52, adjusted by covariates.

	CASC ($N = 285$)	Control ($N = 287$)	*P*- value[a]
Change from weeks 0 to 24			
MPSS urge total	−1.87 [−2.01 to −1.72]	−1.73 [−1.88 to −1.59]	0.009
MPSS total excluding urges	−0.59 [−0.72 to −0.47]	−0.42 [−0.56 to −0.29]	<0.001
FTCQ-12 emotionality	−1.70 [−1.89 to −1.51]	−1.32 [−1.49 to −1.14]	<0.001
FTCQ-12 expectancy	−2.46 [−2.68 to −2.25]	−2.16 [−2.37 to −1.95]	0.002
FTCQ-12 compulsivity	−1.74 [−1.94 to −1.54]	−1.74 [−1.93 to −1.55]	0.202
FTCQ-12 purposefulness	−2.84 [−3.10 to −2.58]	−2.17 [−2.44 to −1.90]	<0.001
FTCQ-12 general craving score	−2.09 [−2.25 to −1.93]	−1.78 [−1.93 to −1.63]	<0.001
KTSND	−7.0 [−7.7 to −6.2]	−3.9 [−4.5 to −3.2]	<0.001
Change from weeks 0 to 52			
MPSS urge total	−1.82 [−1.98 to −1.67]	−1.65 [−1.81 to −1.49]	0.007
MPSS total excluding urges	−0.52 [−0.64 to −0.39]	−0.42 [−0.54 to −0.29]	0.061
FTCQ-12 emotionality	−1.60 [−1.80 to −1.41]	−1.21 [−1.39 to −1.03]	0.001
FTCQ-12 expectancy	−2.39 [−2.60 to −2.19]	−2·10 [−2.33 to −1.88]	0.002
FTCQ-12 compulsivity	−1.71 [−1.90 to −1.52]	−1.55 [−1.75 to −1.35]	0.019
FTCQ-12 purposefulness	−2.84 [−3.10 to −2.58]	−2.00 [−2.28 to −1.73]	<0.001
FTCQ-12 general craving score	−2.03 [−2.19 to −1.87]	−1.65 [−1.81 to −1.48]	<0.001
KTSND			

Mean [95% CI] scores are p
MPSS Mood and Physical Sy
Kano Test for Social Nicotin
[a]Analysis was based on ana

24 weeks that obtained an

Article | Open Access | Published: 12 March 2020

A randomized controlled trial of a smoking cessation smartphone application with a carbon monoxide checker

Katsunori Masaki, Hiroki Tateno ✉, Akihiro Nomura, Tomoyasu Muto, Shin Suzuki, Kohta Satake, Eisuke Hida & Koichi Fukunaga

npj Digital Medicine **3**, Article number: 35 (2020) | Cite this article

Ontologies provide ‘feature types’ for annotations

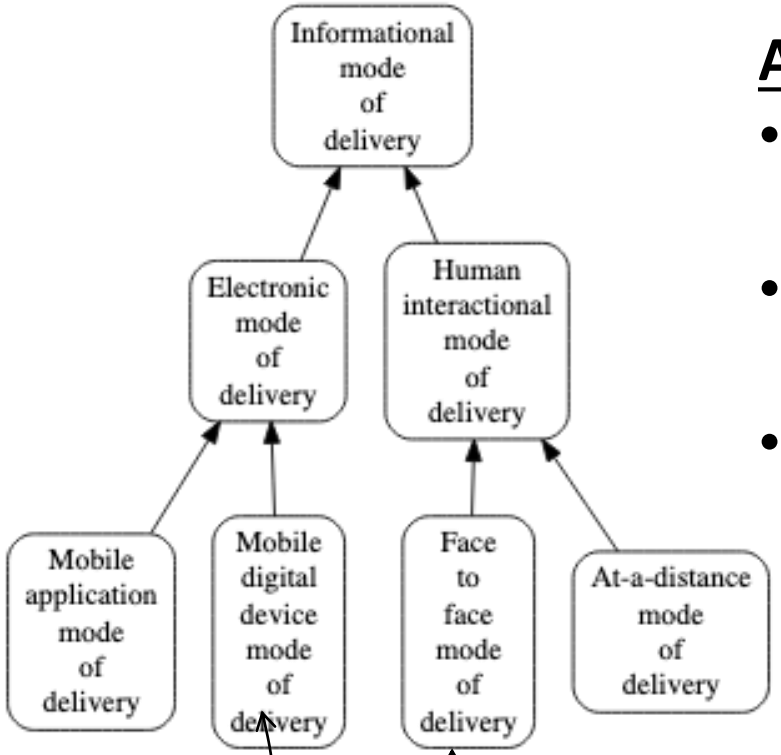
Abstract

Background: Tobacco is a major public health concern. A 12-week standard smoking cessation program is available in Japan; however, it requires face-to-face clinic visits, which has been one of the key obstacles to completing the program, leading to a low smoking cessation success rate. Telemedicine using internet-based video counseling instead of regular clinic visits could address this obstacle.

Objective: This study aimed to evaluate the efficacy and feasibility of an internet-based remote smoking cessation support program compared with the standard face-to-face clinical visit program among patients with nicotine dependence.

Methods: This study was a randomized, controlled, open-label, multicenter, noninferiority trial. We recruited nicotine-dependent adults from March to June 2018. Participants randomized to the telemedicine arm received internet-based video counseling, whereas control participants received standard face-to-face clinic visits at each time point in the smoking cessation program. Both arms received a CureApp Smoking Cessation smartphone app with a mobile exhaled carbon monoxide checker. The primary outcome was a continuous abstinence rate (CAR) from weeks 9 to 12. Full analysis set was used for data analysis.

Results: We randomized 115 participants with nicotine dependence: 58 were allocated to the telemedicine (internet-based video counseling) arm and 57, to the control (standard face-to-face clinical visit) arm. We analyzed all 115 participants for the primary outcome. Both telemedicine and



Annotation types:

- Presence/Absence (e.g. interventions)
- Categorical (e.g. country)
- Quantitative (e.g. mean age)

Annotations consist of feature, value, context

Age	46 ± 11	47 ± 11	45 ± 12
Age ranges			
<40 years	171 (30)	75 (26)	96 (33)
40–49 years	179 (31)	97 (34)	82 (29)
50–59 years	159 (28)	82 (29)	77 (27)
≥60 years	63 (11)	31 (11)	32 (11)
Male sex	426 (75)	216 (76)	210 (73)
Smoking history			
Years of smoking	25 (18–33)	25 (19–32)	24 (17–33)
Cigarettes per day	20 (15–20)	20 (15–20)	20 (15–20)
Pack-years	20 (14–30)	21 (14–30)	20 (13–31)
Exhaled CO (ppm)	17 ± 11	17 ± 10	18 ± 11
TDS score	7.7 ± 1.5	7.7 ± 1.4	7.8 ± 1.5
FTND	5.3 ± 2.1	5.2 ± 2.0	5.3 ± 2.1
Comorbidities			
Cardiovascular diseases	93 (16)	45 (16)	48 (17)
Respiratory diseases	93 (16)	48 (17)	45 (16)
Psychiatric diseases	31 (5)	12 (4)	19 (7)
Prescribed Medication			
Varenicline	454 (79)	227 (80)	227 (79)
Nicotine patch	114 (20)	56 (20)	58 (20)
No medication	4 (1)	2 (1)	2 (1)

Data include mean ± standard deviation, number (%), or median (interquartile range) scores.
CO carbon monoxide, *FTND* Fagerström test for nicotine dependence, and *TDS* tobacco dependence screener.

novel smartphone system for smoking cessation integrated with a

cessation may be a promising strategy to reduce smoking prevalence worldwide, and future research on a global scale is warranted.

METHODS

Study design

This was a multi-center, randomized, controlled, open-label trial (Fig. 2). A detailed study protocol was presented elsewhere[28]. Briefly, participants were recruited from 31 smoking cessation clinics in Japan, from October 2017 to January 2018, and allocated 1:1 to the CASC intervention group and the control group. The intervention group used the CASC system, and the control group used the control app, each for 24 weeks, in addition to a 12-week standard smoking cessation treatment. In reference to the varenicline maintenance therapy for 24 weeks[22], we were interested in whether or not the effectiveness of the CASC system could be maintained after discontinuation of using the app, and we limited access to the app to 24 weeks and evaluated CARs up to 52 weeks. The study was performed in compliance with CONSORT statements. The protocol and written informed consent forms were reviewed and approved by the Institutional Review Board at Keio University School of Medicine and all affiliated institutions. This trial was registered at the University Hospital Medical Information Network (UMIN) Clinical Trials Registry (UMIN000031589).

Participants

We recruited nicotine-dependent adults who visited outpatient clinics to receive smoking cessation treatment under the Japanese National Health Insurance program. Physicians at each clinic screened the patients and obtained written informed consent. The baseline data were collected using a self-administered questionnaire at the initial visit. The program consisted of five visits during a 12-week period, and counseling and pharmacotherapy, including nicotine patch or varenicline were provided by physicians[6]. Inclusion criteria were as follows: adult current smokers who (1) were diagnosed with nicotine dependence (tobacco Dependence Screener [TDS] score ≥5 points)[29]; (2) had a smoking history of pack-years ≥10; (3)

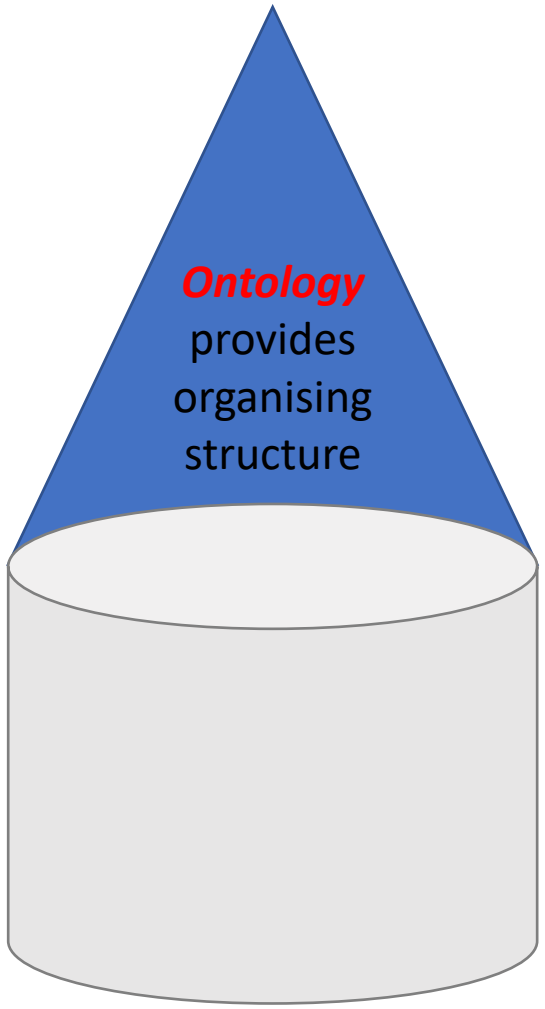
Feature:
Mean number of times tobacco used
Specific value:
'20'
Surrounding text:
Cigarettes per day 20 (15-20)

Feature:
varenicline
Specific value:
'varenicline maintenance therapy for 20 weeks'
Surrounding text:
'In reference to the varenicline maintenance therapy for 24 weeks, …'

BCI database

manually annotated
'gold standard'

evidence 'surveillance' and
automated feature extraction

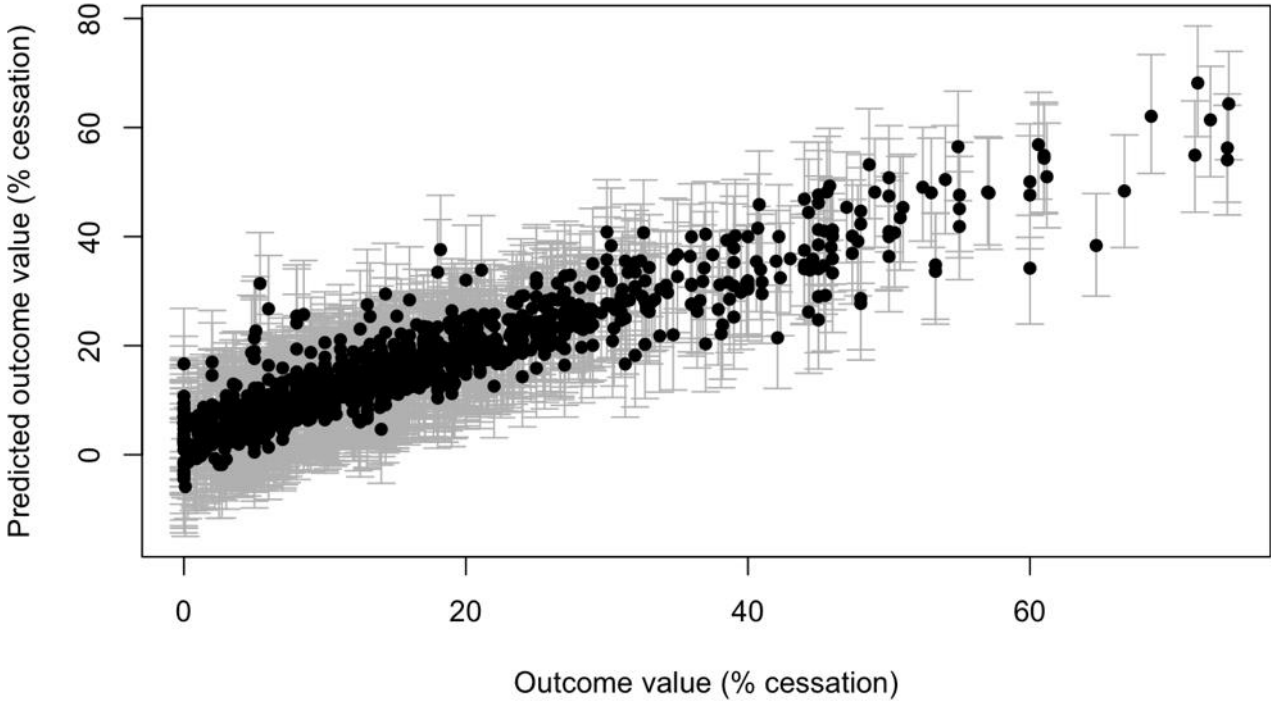


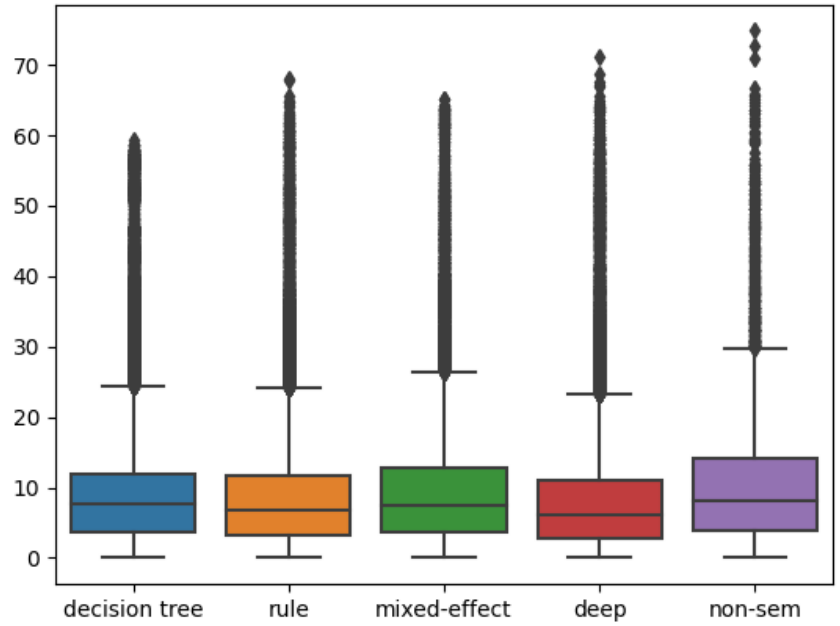
Ontology:
734 classes

Smoking cessation:
455 papers, 1098 arms
55 attributes

Physical activity:
110 papers, 241 arms
160 attributes

Predicting smoking cessation intervention outcomes is hardly an exact science



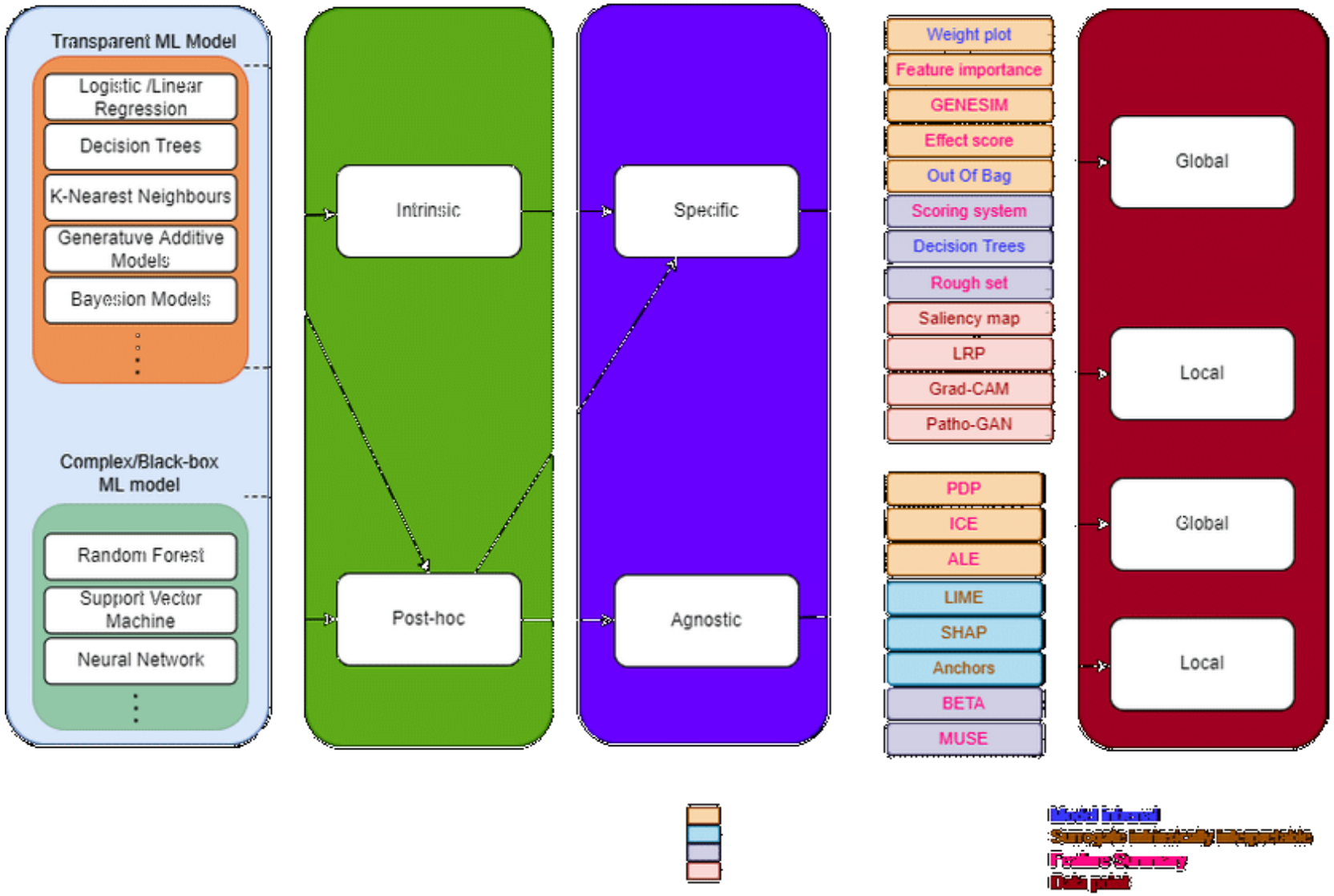


MAE from train/test split (80:20 x 100)

For policy-makers, what is most important is the **explanation** of the prediction

Recommendations are highly dependent on tradeoffs amongst parameters and constraints when searching for optimal strategies, see, e.g.

Xu et al., Statistical Methods in Medical Research, 2020 29 (11) p 3113-3134



Transparent ('glass-box') ML models are fully interpretable

Abdullah et al., **A Review of Interpretable ML in Healthcare: Taxonomy, Applications, Challenges, and Future Directions.** December 2021. Symmetry 13(12):2439

Rules are the most interpretable ML approach

Rules have the form:

if *antecedent* **then** *consequent*

E.g.

if *(smoking)* **then** *(lung cancer risk)*

Antecedent can be a logical combination of features:

if *A* **and** *B* **and** *C* **and** … **then** *consequent*

E.g. **if** *(sitting all day)* **and** *(age>30)* **then** *(gaining weight)*

Pure rules-based systems are too brittle and not trainable enough for sufficient performance

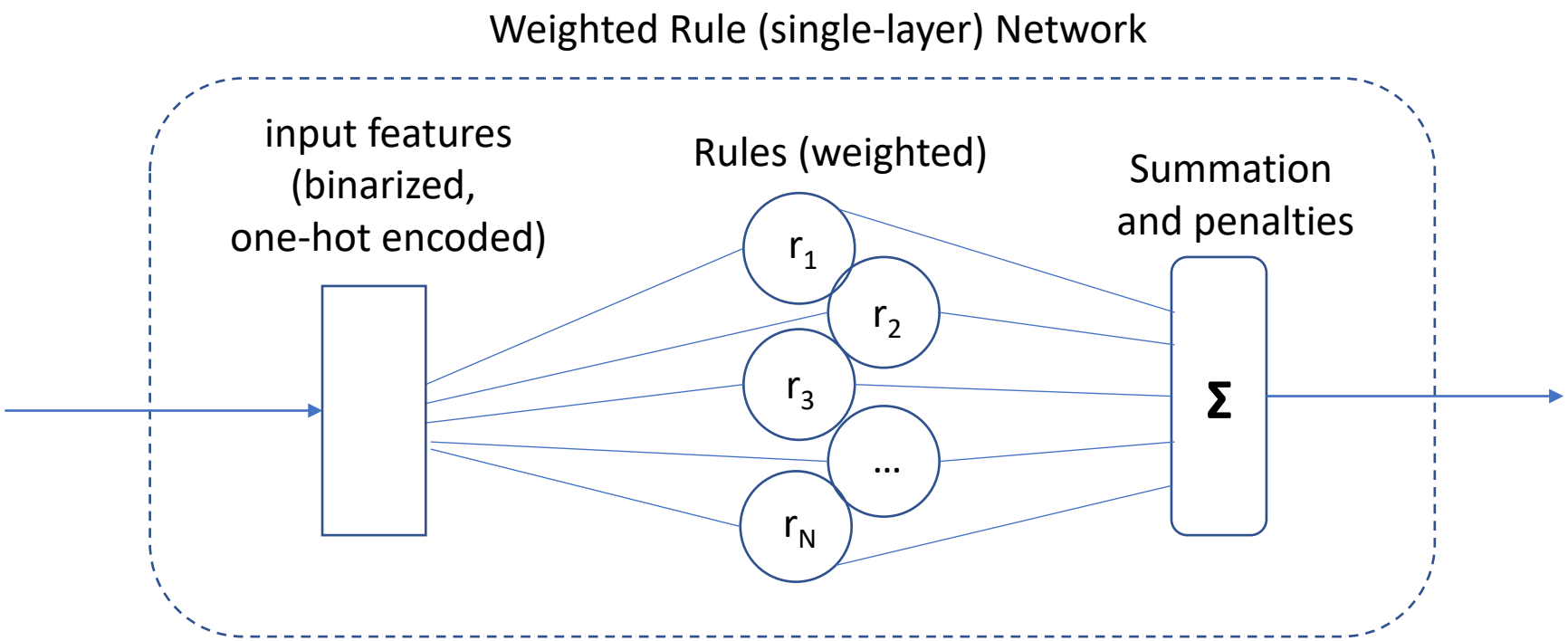
Advantages

- Naturally interpretable *(as long as rules are short and there are not too many of them)*
- Can in principle be induced from data *(but see disadvantages)*
- Easy integration with semantics, logical structures and constraints

Disadvantages

- (usually) only apply for categorical prediction problems
- Require binary features as inputs
- Not differentiable, so not trainable using efficient learning approaches
- Prone to overfitting
- Too rigid to cope with uncertainties

A ‘best of both worlds’ hybrid approach: trainable additive weighted rule sets with semantic penalties *



predicted outcome:
a sum of
(the weighted ‘fit’
of each rule multiplied
by the base rule weight)

and ‘most applicable’ rules

$a_{1,1}F_1$ & $a_{1,2}F_2$ & … & $a_{1,n}F_n$ & $b_{1,1}$(not F_1) & $b_{1,2}$ (not F_2) & … & $b_{1,n}$(not F_n) → (+/-) r_1

** Glauer and Hastings, in preparation*
-- HBCP/Semantic-Prediction on GitHub

A ‘best of both worlds’ hybrid approach: trainable additive weighted rule sets with semantic penalties *

Disadvantages of rules systems

- (usually) only apply for categorical prediction problems

Each rule is assigned a ‘fit’ and a ‘weight’
Rules within sets added together to get the prediction

- Require binary features as inputs

We binarise using fuzzy strategies

- Not differentiable, so not trainable using efficient learning approaches

Weights are trainable using backpropagation

- Prone to overfitting

We introduce various semantic penalties

- Too rigid to cope with uncertainties

Fuzzy interpretations allow generalisation

** Glauer and Hastings, in preparation*
-- HBCP/Semantic-Prediction on GitHub

Input features pre-treatment, binarization

- Logical implication is pre-reasoned in feature dataset
  (e.g. if ‘nurse’ or ‘doctor’ then ‘healthcare professional’)
- Binary features are retained, categorical features are one-hot encoded, and continuous features are binarized with fuzzy boundaries using several strategies as semantically appropriate:
  - Meaningful semantic categories, e.g. age = child, young adult, older adult, …
  - Fixed-width categories, e.g. #times.tobacco.smoked <5, <10, <15, …
  - Fixed-quantile categories, e.g. proportion.female …

    (these numeric category boundaries are fuzzy)
  - Fixed (categorical, not fuzzy) numeric values e.g. 100% female

** Glauer and Hastings, in preparation*

Semantic penalties – what makes ‘good’, interpretable rules?

- Prefer
  - shorter rules
  - combinations of features that cross semantic domains (e.g. intervention type, setting, population)

- Avoid (within a single rule)
  - Mutually disjoint classes, unless belonging to different semantic domains
  - Hierarchically related classes
  - F and (not F), for any feature F
  - Manually specified excluded combinations e.g. varenicline and (not pharmacological support)

** Glauer and Hastings, in preparation*

Results

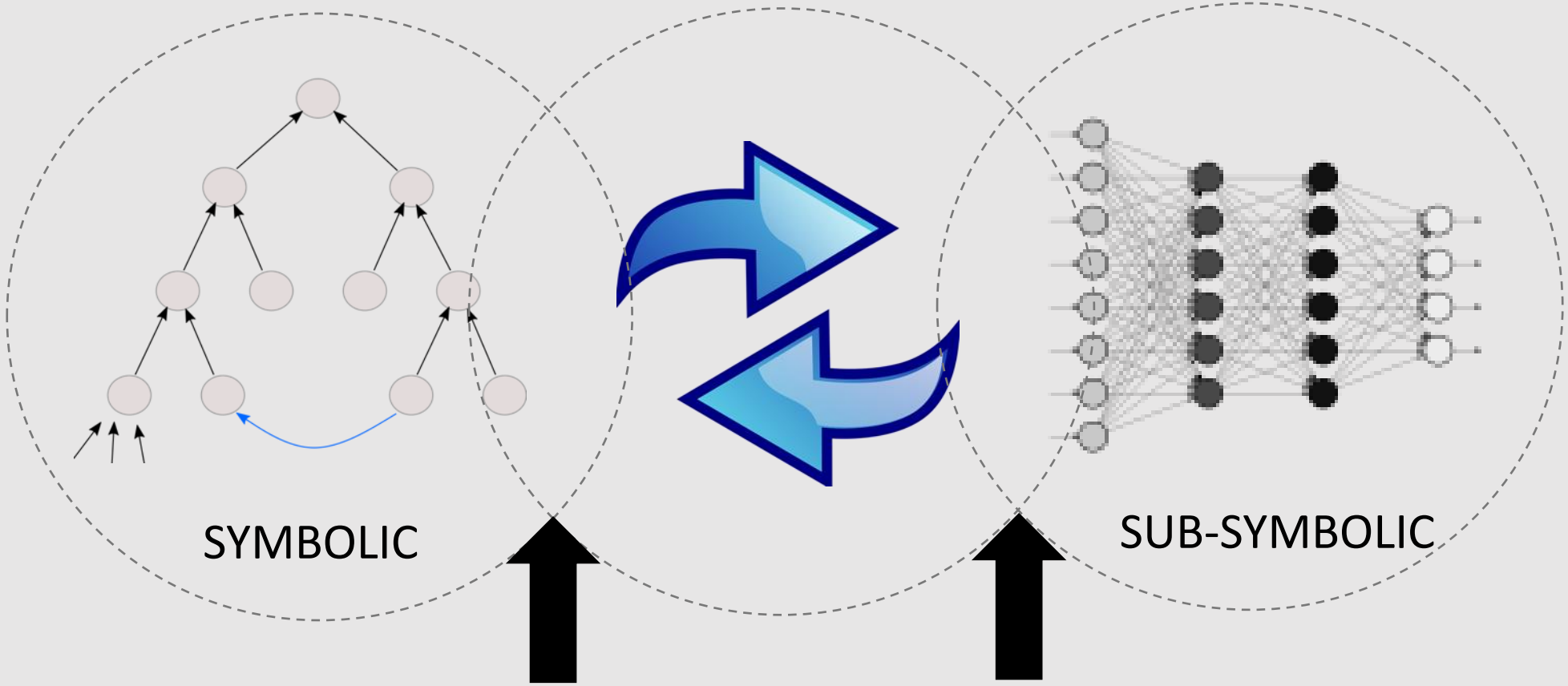
- All variants have the same accuracy (MAE ~8, comparable to DL)
- No penalties: 100 rules x 300 rule parameters, **not interpretable**
- With penalties for length and crispness of rules: ~ 20 rules x ~ 10 parameters, **somewhat interpretable**
- With semantic penalties: ~ 15 rules x ~ 1-5 parameters, **interpretable**

Summary

- For smaller datasets where interpretability matters, trainable rules are a very good alternative with comparable performance to deep models, and can be used both for classification and regression problems

- Domain knowledge in the form of categories, hierarchy, implications and disjoints can be used to supplement the trained rules with semantic penalties that force the system to learn rules that ‘make sense’

Two case studies: (1) evidence synthesis and (2) (automated) semantic annotation of chemicals



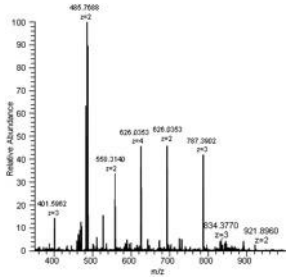
Interpretable prediction of the outcomes of smoking cessation interventions

Interpretable semantic annotation of chemical structures to a chemical ontology

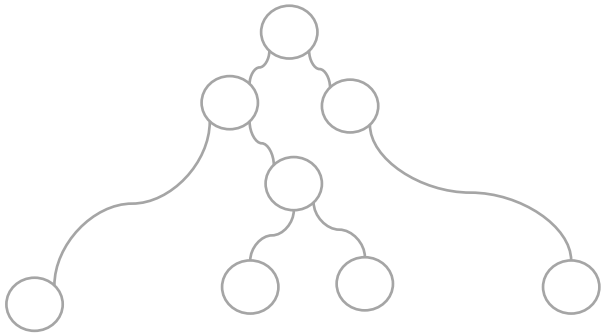
Automating Semantic Annotation in Chemistry

Predicting Chemical Ontology Classes for Chemical Structures

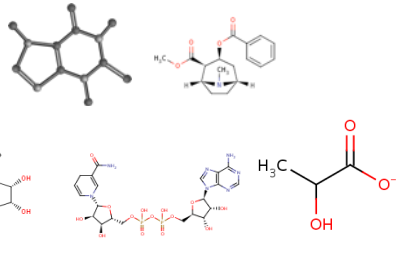
Metabolism is a key differentiator between health and disease, but data can be difficult to interpret

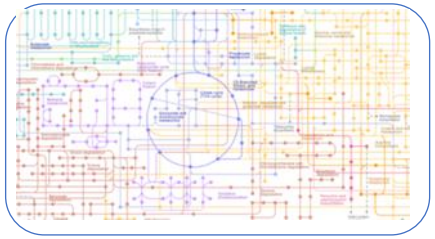


ChEBI

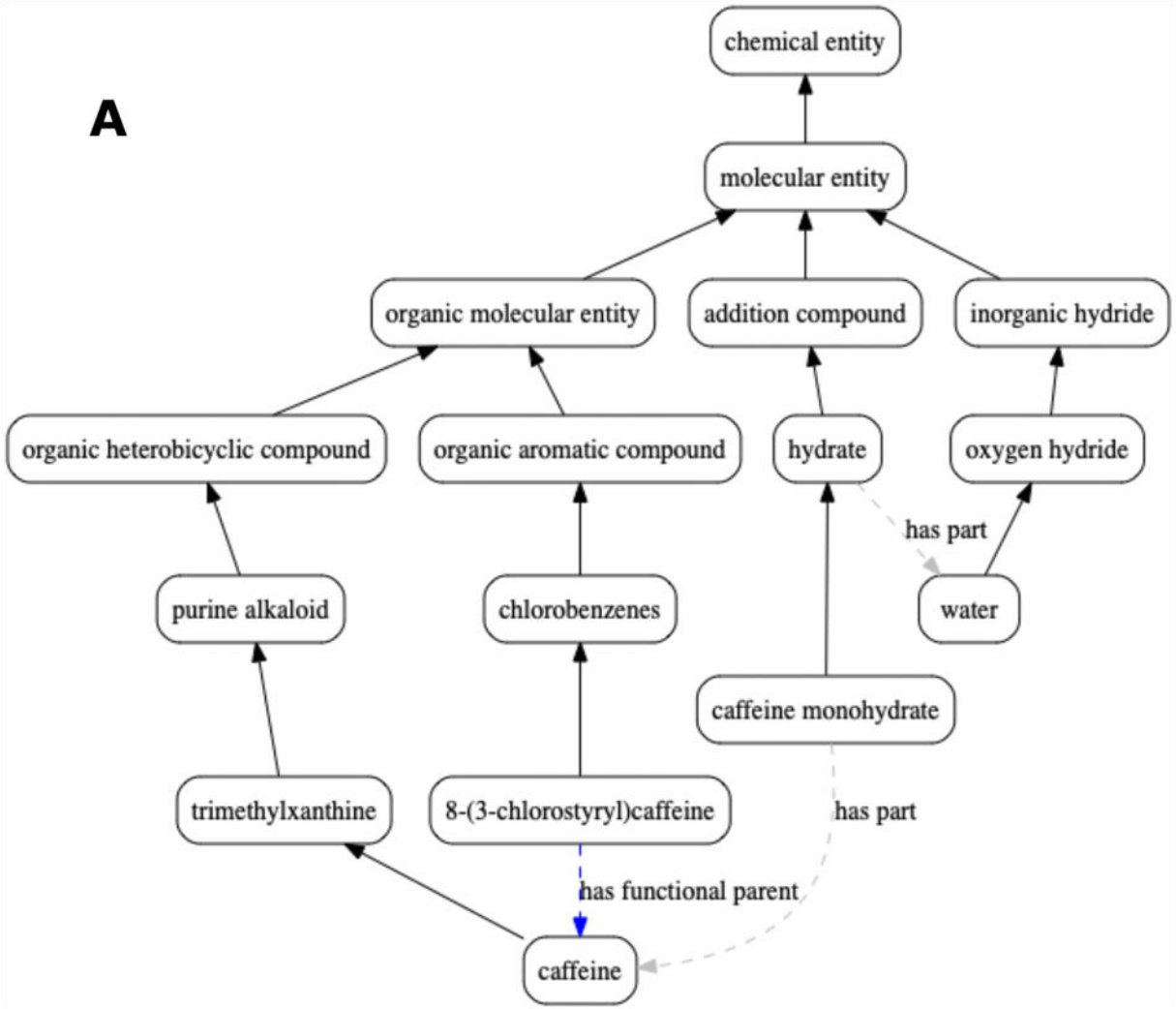


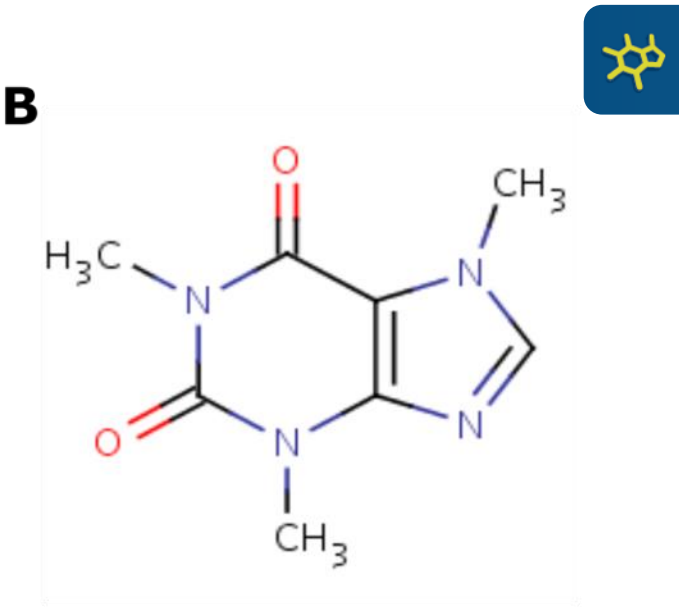
Metabolome





Chemical ontologies provide a link between chemical nomenclature, structures and annotations



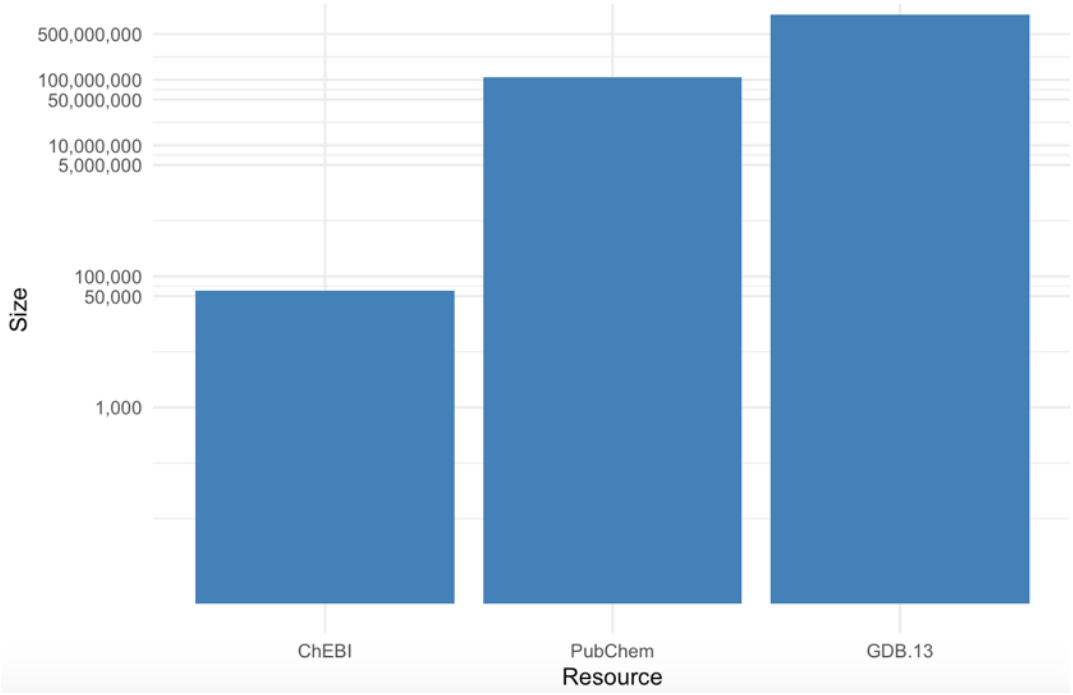


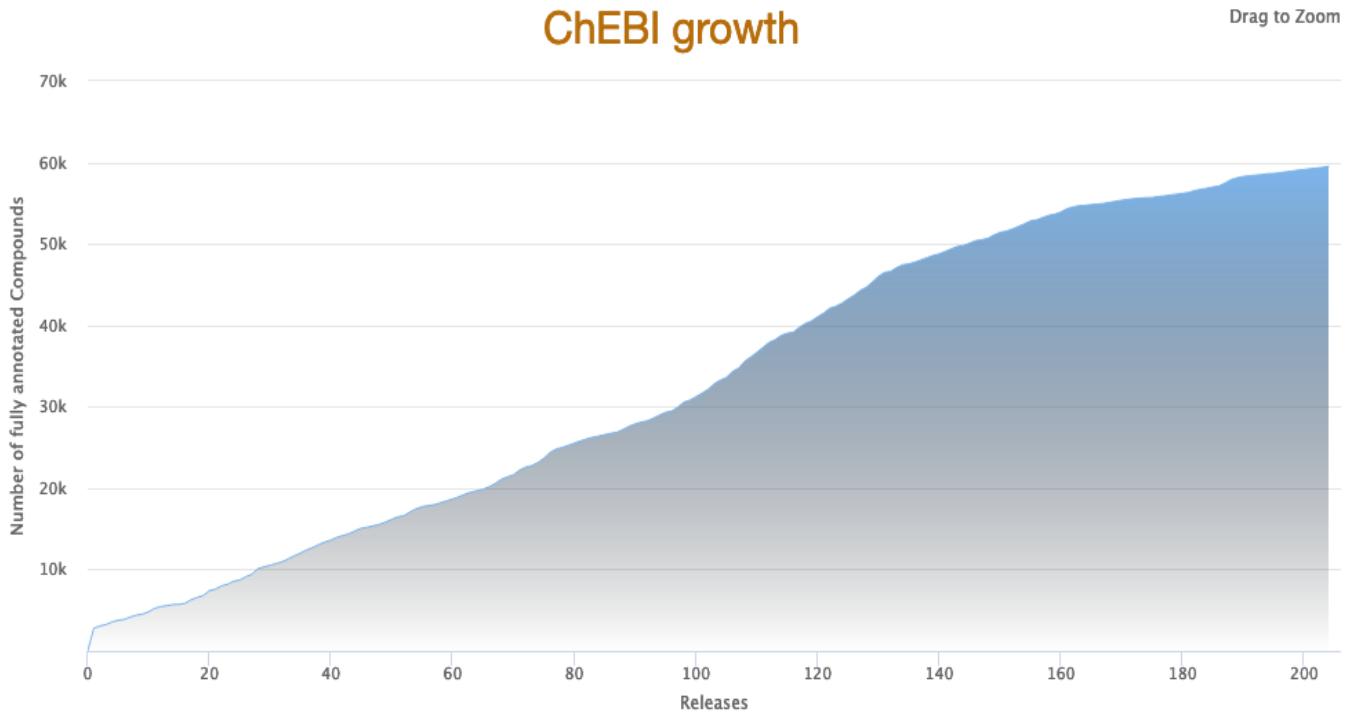
ChEBI

ChEBI Name	**caffeine**
ChEBI ID	**CHEBI:27732**
SMILES	Cn1cnc2n(C)c(=O)n(C)c(=O)c12
Formula	C8H10N4O2
Net Charge	0
Average Mass	194.19076

Hastings et al. "**Learning Chemistry: Evaluating machine learning for the task of structure-based chemical ontology classification,**" *Journal of Cheminformatics* 2021

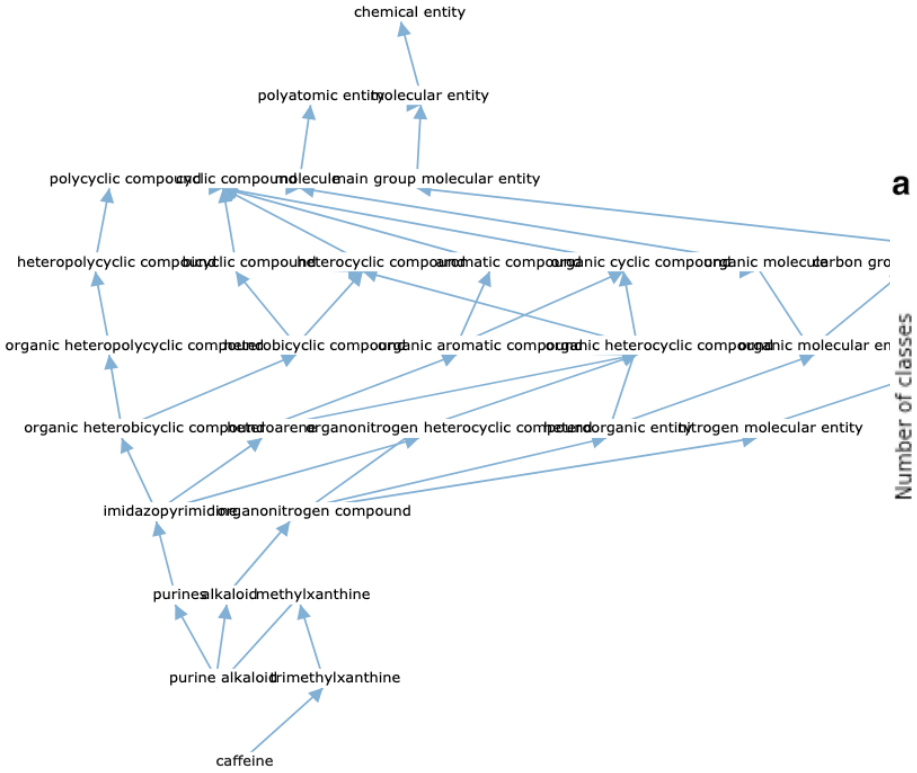
Manually maintained ontologies and knowledge resources have high quality, but poor scalability



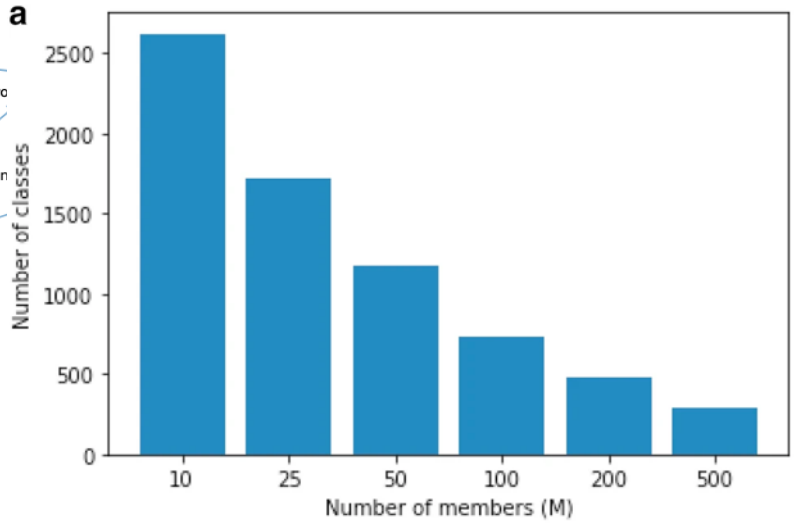


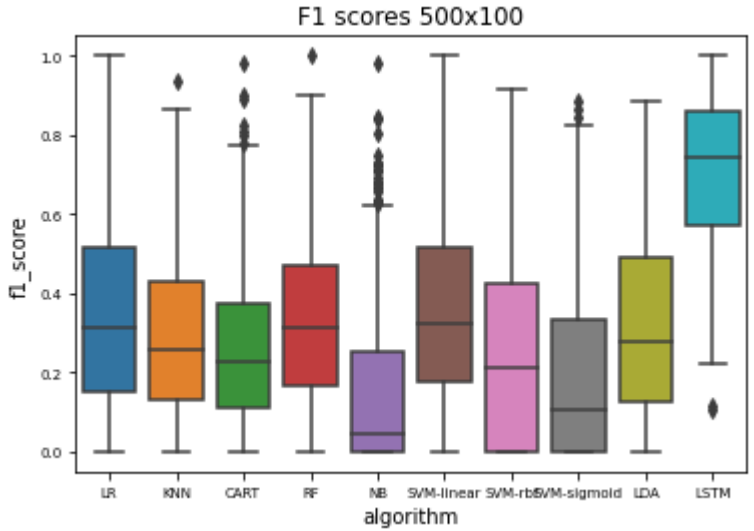
Hastings et al. "**Learning Chemistry: Evaluating machine learning for the task of structure-based chemical ontology classification,**" *Journal of Cheminformatics* 2021

The shape of the ontology hinders learning, and requires a ‘deep’ neural network



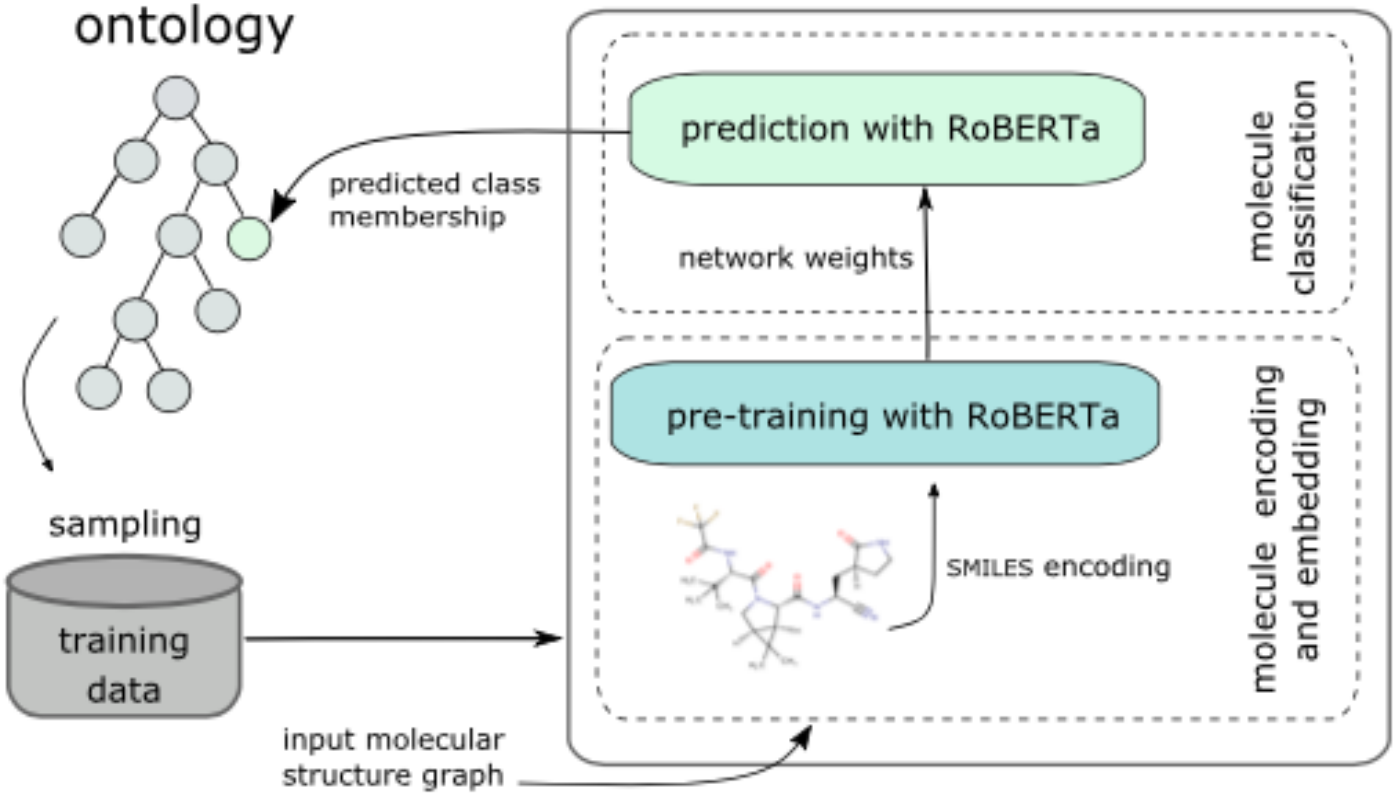
“Extreme” multi-class, polyhierarchical, sparse, very unbalanced





Hastings et al. “**Learning Chemistry: Evaluating machine learning for the task of structure-based chemical ontology classification,**” *Journal of Cheminformatics* 2021

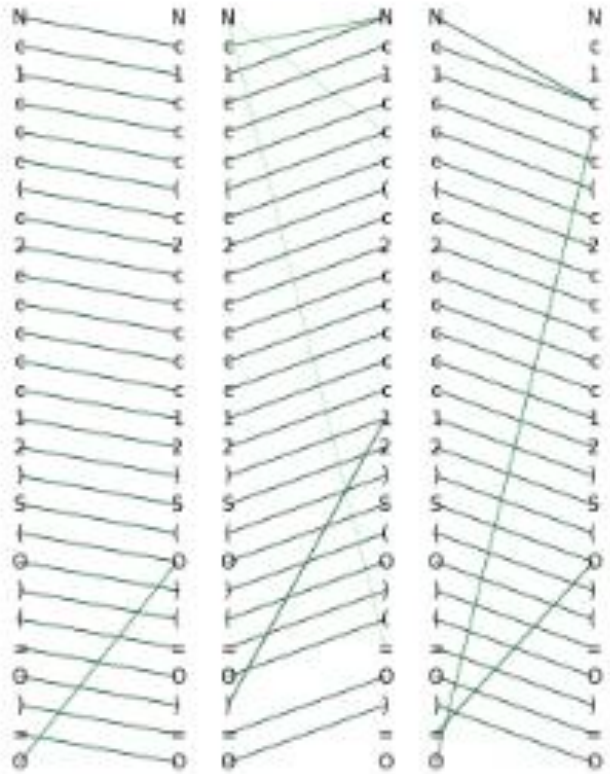
Transformer-based models with pre-training give best performance for this task



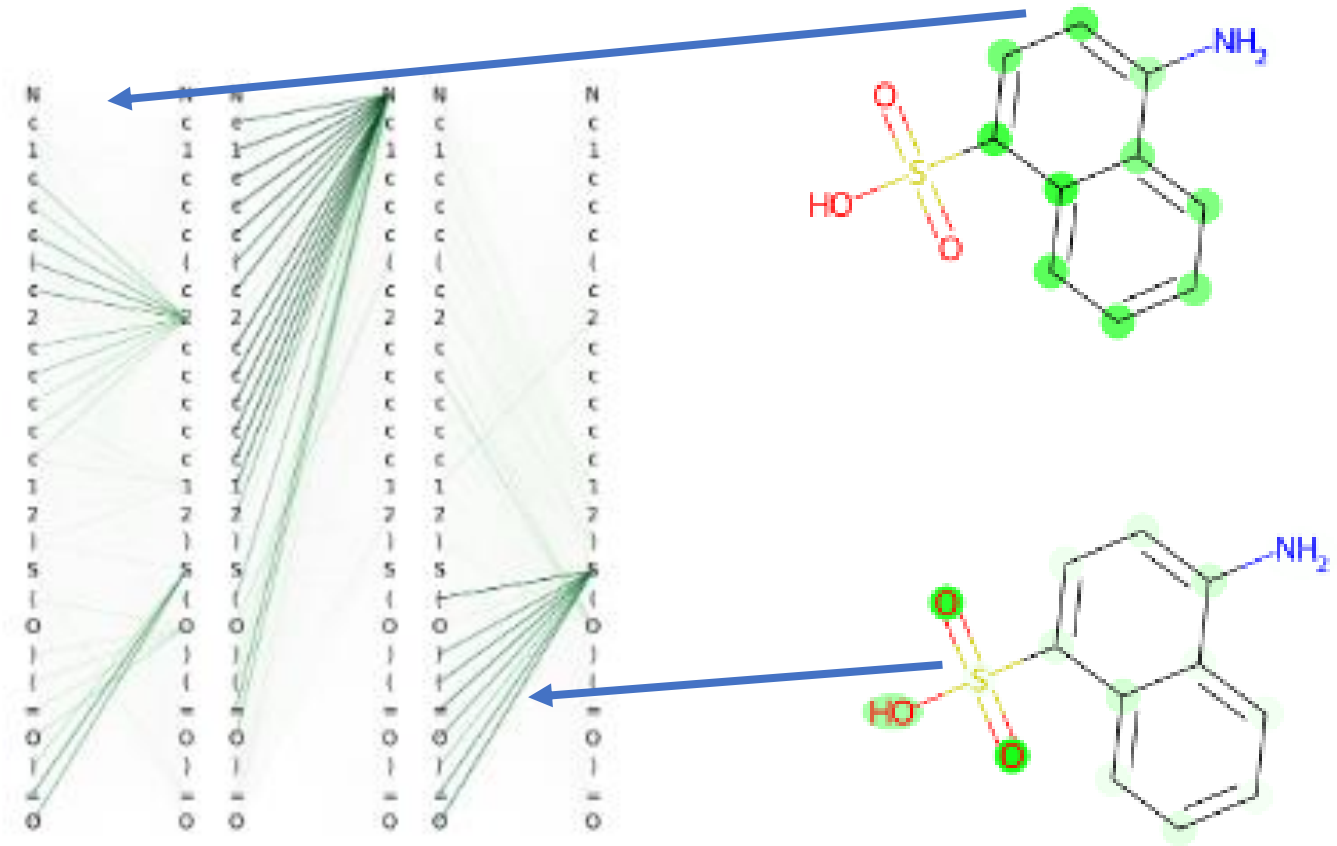
* now we are using Electra

Memariani et al., Proc. of DAO-XAI 2021

The network trained on this semantic task has (somewhat) interpretable attention weights



(a) Attention of naphthionic acid (CHEBI:38219) in layer 2, heads 1-3.

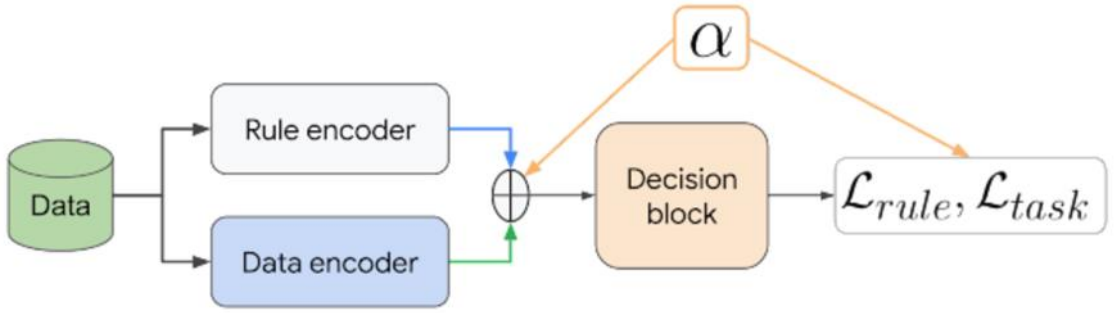


(b) Attention of naphthionic acid (CHEBI:38219) in layer 5, heads 4-6.

Glauer et al. *Interpretable ontology extension in chemistry* SWJ 2022

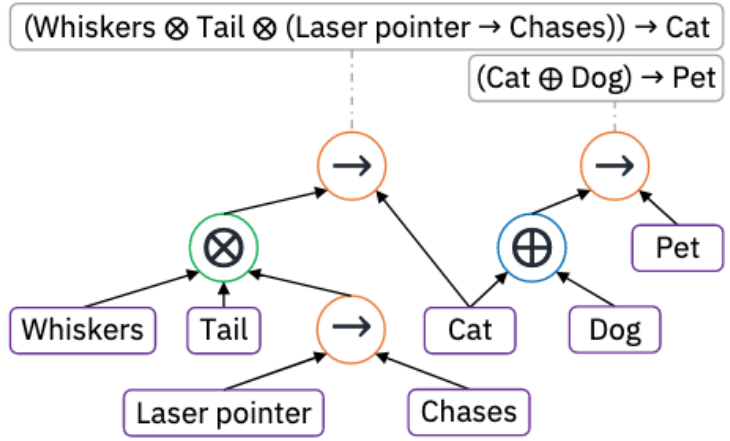
Logical rules can be integrated with DNNs to improve predictive performance

Our current approach: post-hoc correction based on **ontology-driven explicit disjoints**



DeepCTRL pairs a data encoder and rule encoder, which produce two latent representations, which are coupled with corresponding objectives. The control parameter α is adjustable at inference to control the relative weight of each encoder.

DeepCTRL – Seo et al. NeurIPS 2021 https://arxiv.org/pdf/2106.07804.pdf



(a) The LNN graph structure reflects the formulae it represents.

Logical Neural Networks – Riegel et al. NeurIPS 2020 https://arxiv.org/pdf/2006.13155.pdf

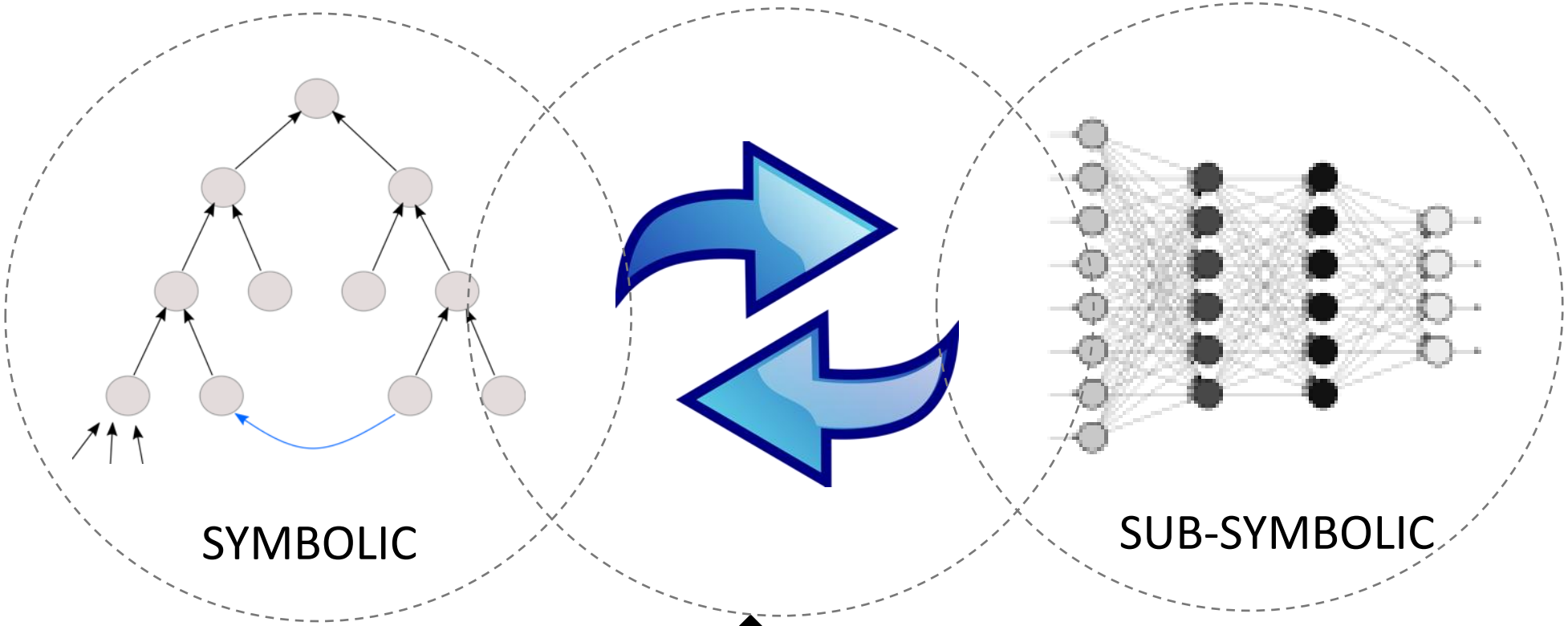
Summary

- For problems that require ‘black box’ deep neural networks, consideration of semantics can nevertheless improve performance and interpretability

- Training a transformer-based model on a semantic prediction task leads to attention weights that correspond to semantic features

- Logical rules and constraints can be incorporated into network training via hybrid architectures with semantic objective functions

Conclusion: ‘best of both worlds’ architectures offer a good option for many problems in medicine



Rigid, True or False, Based on Logic
+ Provable outcomes, Transparent

Quantitative, Associative, Based on Statistics
+ Scalable, Robust to inconsistencies

Data-centric AI / Neuro-symbolic AI / Semantic AI / ‘Informed’ ML / ‘Broad’ AI

Acknowledgements

UCL / HBCP

- Susan Michie
- Robert West
- Alison Wright
- James Thomas
- Emma Norris, Ailbhe Finnerty Mutlu, Paulina Schenk, **and many others**

OvGU Magdeburg

- Martin Glauer
- Fabian Neuhaus
- Adel Memariani
- Till Mossakowski

Thank you! Questions?