Neurosymbolic AI in Medicine: Challenges and Opportunities

Ernesto Jiménez-Ruiz* Mehwish Alam Roberto Barile Jiaoyan Chen
Michael Cochez Roberto Confalonieri Natalia Díaz-Rodríguez
Claudia d'Amato Artur d'Avila Garcez Eleonora Giunchiglia
Dagmar Gromann Sven Hertling Pascal Hitzler Franklyn Howe
Egor V. Kostylev Alessandra Mileo Lia Morra Raghava Mutharaju
Heiko Paulheim Catia Pesquita Rita T. Sousa Valentina Tamma
Annette ten Teije Son N. Tran Janna Hastings†

March 12, 2026

Abstract

Artificial intelligence has many applications in medicine, including drug discovery, diagnostics, clinical decision support, and the automation of clinical documentation. Recently, large language model-based systems such as ChatGPT have made headlines for their groundbreaking performance on a wide range of tasks. However, these systems can make errors, be biased, or act harmfully in certain contexts. New approaches are needed which can complement existing approaches while addressing safety and responsibility. Neurosymbolic AI is a novel paradigm that offers methods that are able to learn from data but remain transparent and controllable. This is achieved by combining data-driven approaches with knowledge-driven approaches.

1 Introduction

Medicine has seen an explosion of recent applications of artificial intelligence (AI), with the potential to make a transformative impact on health and disease outcomes [CKM+23]. Key application areas include the automated management of medical evidence (*e.g.*, [DZP+24]), drug discovery (*e.g.*, [WZV+23]), medical education (*e.g.*, [AAAA+23]) and clinical decision support (*e.g.*, [BWS+23]). Recently, two Nobel Prizes were awarded for the development of key AI technologies that are already having far-reaching effects in medical research [LG24].

However, applications of AI in medicine pose a high safety and regulatory challenge, and there are significant potential risks associated with indiscriminate deployment [BP24]. Key concerns include *bias* in training data, which can lead to unequal or unsafe outcomes across patient groups [Has24]; *unreliability* in model predictions, particularly when systems are applied beyond their validated contexts [BPP+26]; and *uncertifiability*, which limits their use in critical clinical environments where explainability and verification are essential [DHW+25]. From a personalised medicine perspective, AI models often struggle with the "long tail" of rare conditions and patient variability, raising further questions about their generalisability. Consequently, model actionability—the ability of AI systems to produce trustworthy, interpretable, and clinically relevant recommendations—remains central to their effective and ethical integration into medical decision-making [EJG+23].

Neurosymbolic AI [dGL23] is a key technological innovation for medicine, as it combines the strengths of neural networks and symbolic AI, two technologies that, independently, have successfully been applied in medicine (*e.g.*, analysis of MRI scans with computer vision - a neural apporach; use of SNOMED CT to harmonise codes across NHS

*City St George's, University of London, UK. ernesto.jimenez-ruiz@citystgeorges.ac.uk
†Idiap Research Institute, Switzerland. janna.hastings@idiap.ch

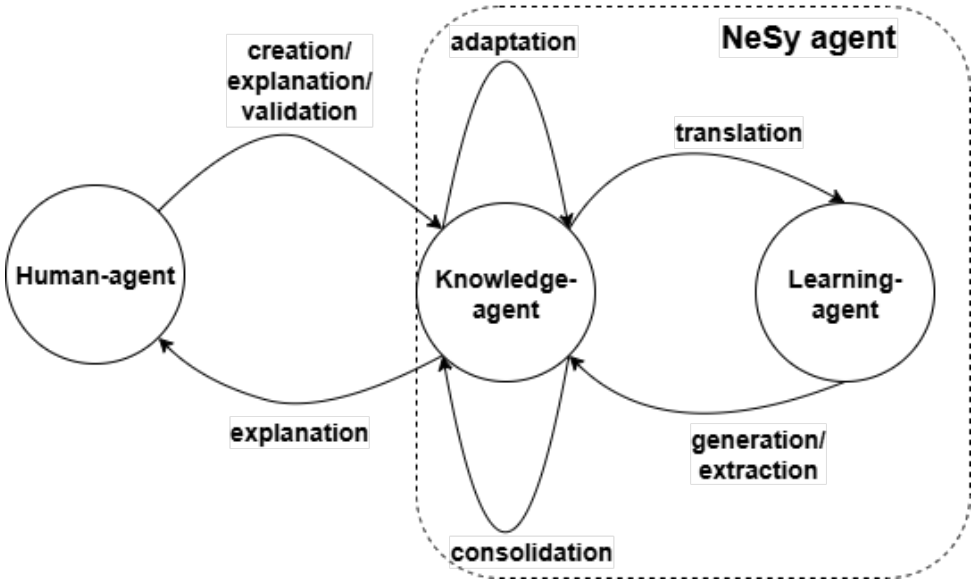

Figure 1: Neurosymbolic AI agent architecture for medicine [HSS+25]

systems - a symbolic approach). Each technological paradigm has weaknesses in isolation. For example, deep learning models lack explanations of their predictions, a key component for an AI system in medicine, and cannot be verified as to how they will perform in new contexts, while symbolic systems are transparent and verifiable but are hard to develop and maintain and struggle to handle unstructured inputs. Neurosymbolic AI systems aim at overcoming these limitations, as they can both learn and reason from (multimodal) data (images, audio, text, video, tables) and have transparency and verifiability implemented through explicit (domain) knowledge.

With the rapid acceleration of the capabilities of LLMs for medical applications, clinicians and stakeholders are looking for ways to implement this key technological innovation safely in clinics [DHP+25]. Neurosymbolic approaches offer a compelling solution to address this need by augmenting language model and related technologies with transparency and guardrails against bias and other problems.

In the next section, we give an introduction to neurosymbolic approaches to AI and highlight what they can offer for medicine. After that, we will discuss some recent success stories for neurosymbolic approaches to AI in medicine and finally close with some remaining open questions.

2 What is Neurosymbolic AI?

Neurosymbolic (NeSy) AI is about the integration of learning and reasoning, two fundamental but traditionally separate areas of AI. There are different ways and specific approaches for combining learning and reasoning; however, although there are different proposals (*e.g.*, [BH05, SZEH21, dGL23, vHtT19]), defining what constitutes a NeSy system remains an open challenge. A first question concerns the learning component: must it be neural, or can symbolic approaches such as inductive logic programming [CD22] also fall within the NeSy paradigm? A second issue relates to the nature of reasoning: must it be symbolic in a strict logical sense, or do differentiable reasoning methods (*e.g.*, [WDWK19]) also qualify? Likewise, should large language models (LLMs)—which exhibit emergent reasoning capabilities over symbolic tokens—be considered inherently neurosymbolic, and are deep learning systems that perform symbolic tasks necessarily so? A third question concerns the degree of integration between learning and reasoning. How tightly coupled must these components be—for example, does a graph-based retrieval-augmented

generative AI system linking an LLM with a knowledge graph (*e.g.*, [GLH+25]) qualify as NeSy? Finally, the definition of a "symbol" itself remains elusive, blurring the line between symbolic and sub-symbolic representations [HSS+25].

In the process of combining learning and reasoning, a few challenges of a computational nature arise: learning from data, with state-of-the-art neural networks, uses distributed vector representations (embeddings) typically in continuous space, while formal reasoning normally takes place using discrete symbol manipulation to prove a conclusion from existing knowledge. To illustrate the difference in the context of medical diagnosis, a state-of-the-art medical imaging AI system (*the learning agent*) would take vector representations of an X-ray or MRI scan and use a neural network to learn to distinguish whether an image is diseased or healthy (*e.g.*, whether a chest X-ray indicates the occurrence of pleural effusion). A symbolic rule, managed by the *knowledge agent* and associated with the same system, might state that an enlargement of the heart (A) and an obscured diaphragm (B) indicate pleural effusion (C), described symbolically as $(A \text{ and } B) \rightarrow C$. The symbolic description serves not only as a possible explanation for the behaviour of the state-of-the-art neural network system, but offers a window into the system with the possibility of checking whether the system's high performance is indeed justified according to medical best practice [PF25, LBH10, dGZ99]. Reasoning about the neural network at the symbolic level may offer guarantees on safety requirements, and furthermore allows domain experts to intervene in the complex AI system if needed, using the transparent symbolic rules for the purpose of system interaction. In this context, a substantial challenge for NeSy AI is the generation/extraction of simple and relevant symbolic concepts from very large and complex neural network models.

We envision NeSy as a hybrid, agent-based architecture for continuous learning and reasoning, composed of three main agents (see Figure 1): *the human agent*, *the knowledge-based agent*, and *the learning agent* [HSS+25]. The human agent contributes domain knowledge, validates new insights, and interacts in a human-in-the-loop manner to refine and interpret the system's reasoning. The knowledge-based agent manages verified knowledge. It mediates between the human and learning agents—responding to queries through symbolic inference, requesting new knowledge from the learning agent when needed, and ensuring that knowledge additions and revisions remain semantically consistent. The learning agent can generate new knowledge either via symbolic (*e.g.*, Inductive Logic Programming) or neural methods (*e.g.*, through deep learning). Human agents play an important role due to their capabilities that differ from those of AI systems [IHvH+25, VAM24]. Thus, our NeSy framework vision *(i)* includes human expertise 'in-the-loop'; *(ii)* explicitly distinguishes and combines deductive (certain) logical reasoning with non-deductive (uncertain) learning and reasoning approaches; and *(iii)* supports diverse learning paradigms beyond neural networks. The inclusion of domain expertise is paramount in a domain like medicine. Ontology-based knowledge graphs (KGs) [Hit21] may play a pivotal role in this NeSy vision by combining symbolic knowledge representation with formal logical reasoning [HJRW25, HDMN25]. This is particularly relevant in biomedicine, where numerous high-quality ontologies and KGs—such as GO, ChEBI, PPI, SNOMED, UMLS, BioPortal, and Bio2RDF— provide rich, structured sources of domain knowledge [CDH+23].

3 Recent successes of NeSy approaches in medicine

This section showcases the potential of neurosymbolic AI solutions in clinical and preclinical tasks.

3.1 Medical evidence

Medical evidence refers to reliable knowledge from medical research that helps doctors make informed decisions about how to diagnose, treat, and care for patients. Medical evidence, however, is typically *locked* in numerous free-text publications.

Use case (medical evidence from clinical trial publications). Kang et al. [KTK+21] introduce a neurosymbolic solution to enhance the extraction of medical evidence from free-text clinical trial publications.

- *Knowledge component.* Medical evidence represented as propositions (*e.g.*, ⟨ "Group A", "decreasesd", "vascular resistance" ⟩).
- *NeSy learning component.* Medical evidence-informed attention mechanism integrated with BioBERT, a pre-trained biomedical language representation model for biomedical text [LYK+20].
- *Lessons learned.* Improvement over the purely neural solution baseline. Similar performance to state-of-the-art with less data. Creation of reusable symbolic representations of medical evidence.

Use Case (medical evidence for better smoking-cessation outcomes). Glauer et al. [GWMH22] present a neurosymbolic AI system that learns rules for predicting behaviour change intervention outcomes, based on a corpus of annotated clinical trials literature.

- *Knowledge component.* The Behaviour Change Intervention ontology [MWF+20] and the learned rules.
- *NeSy learning component.* A neuro-fuzzy model to learn rules. The model is constrained/informed by ontology-based and user-defined rules.
- *Lessons learned.* Performance similar to black-box deep learning solutions, but providing an explainable solution via rules that are compliant with background domain knowledge.

Use case (AI-assisted clinical trial prescreening). Calaprice-Whitty et al. [CWGS+20] evaluate AI-assisted commercial system Mendel.ai for clinical trial participant prescreening in oncology, comparing it with standard manual screening procedures.

- *Knowledge component.* Formalised eligibility criteria derived from clinical trial protocols, including inclusion and exclusion rules structured to reflect regulatory and medical constraints.
- *NeSy learning component.* An AI system processes electronic health records to identify potential candidates by combining natural language processing and structured data extraction with rule-based eligibility matching against trial criteria.
- *Lessons learned.* AI-assisted prescreening improves efficiency and identification rates compared to standard approaches, while maintaining alignment with protocol-defined constraints. This work highlights the need for medical vocabulary KBs with which to match queries expressed in natural language. In particular, Mendel.ai increases from 24 to 50% the correct identification of patients meeting all eligibility criteria for clinical trials.

3.2 Clinical decision support

Clinical decision support refers to tools that help healthcare professionals make better decisions about patient care by providing relevant information and recommendations.

Use case (clinician intervention in automated X-ray diagnoses). Ngan et al. [NMBP+24] introduced a neurosymbolic approach where compact rules extracted from trained convolutional neural networks (CNNs) using ERIC (Extraction of Relations Inferred from Convolutions [TKI21]) are evaluated w.r.t. medically relevant relationships, enabling domain expert intervention into automated decision making.

- *Knowledge component.* A user interface displaying the reasoning of the extracted concepts from a trained CNN generated by ERIC so that clinicians can review the results and, if necessary, intervene in the set of rules to construct an alternative model that only uses meaningful concepts relevant to the X-ray diagnosis in hand.
- *NeSy learning component.* Closing the NeSy cycle via knowledge extraction followed by translation for CNN fine-tuning, a clustering technique proposed in [NGT22] automates the assignment of CNN-trained kernels to medical concepts defined based on the most frequently activated regions of an entire data set. The approach automates the localization of relevant regions in the X-ray by associating concepts to radiomics features.
- *Lessons learned.* The evaluation of CNNs trained on chest X-ray images, allowing domain experts to intervene in the learning process, indicated that kernel groups within a CNN are not randomly learned, but provide conceptual information essential to the performance of the model. Results showed that it is not guaranteed that kernel heat-maps would reveal domain-relevant regions unless a deliberate attempt was made by human experts to define the regions. The approach should help machine learning practitioners streamline CNN models to reduce the amount of computational resources required for CNN training and updates. In this process, it is important to align trained models with established best practices in medical diagnosis, crucial for the research to translate faster into the real-world.

Use case (explainability for sequential EHR predictions). Doctor XAI [PPP20] is a neurosymbolic explainability technique designed for black-box models operating on sequential, multi-label electronic health record (EHR) data. It explains predictions of models using ontology-guided local surrogate rules derived from patient-specific synthetic neighbourhoods.

- *Knowledge component.* The method uses the ICD-9 medical ontology to compute semantic similarity between diagnosis codes, guide neighbourhood construction, and define perturbations over concept groups sharing a least-common superconcept. Ontology structure informs which diagnoses are grouped, masked, or perturbed when generating synthetic patient histories.
- *NeSy learning component.* Doctor XAI generates ontology-aware synthetic sequences around the patient of interest, labels them using a recurrent neural network (Doctor AI), and trains a symbolic multi-label decision tree as a local surrogate. Temporal encoding/decoding bridges sequential inputs and symbolic surrogate models, enabling rule extraction directly tied to visit timelines.
- *Lessons learned.* Ontology-guided perturbations improve the fidelity of the surrogate model to the black-box compared with non-ontological strategies, producing more stable and clinically meaningful explanations. Experiments on MIMIC-III show higher fidelity (up to 0.89) and clearer rule structures when leveraging ICD-9 knowledge. The approach demonstrates that combining medical ontologies with neural sequence models yields transparent, interpretable explanations suited for clinical auditing and decision support.

Use case (symbolic rule extraction from medical predictors). The TREPAN-reloaded [CWBdPM20, CWBM21] algorithm extracts symbolic decision trees from complex neural models used in medical prediction tasks. It is tailored to settings where the underlying model integrates both data-driven learning and domain knowledge, and where clinicians require transparent, verifiable explanations of model behaviour.

- *Knowledge component.* The method queries a structured medical knowledge model, *e.g.*, ontologies or curated clinical concept hierarchies, to constrain the feature space explored during extraction. Knowledge elements help guide split selection, favour clinically meaningful predicates, and support semantic pruning of irrelevant or redundant conditions during tree generation.

- *NeSy learning component.* The algorithm operates as a query-based extractor over the target neural network: synthetic samples are generated around chosen regions, labelled by the network, and used to train a symbolic decision tree that approximates global behaviour. It incorporates knowledge-informed sampling and knowledge-constrained split selection, producing trees whose structure reflects both the neural model's decision boundary and medically relevant concepts.

- *Lessons learned.* The extracted trees provide interpretable rule-based explanations while maintaining fidelity to the underlying neural model. Knowledge-guided extraction yields feature splits that are more aligned with medical semantics, improving clinician interpretability and reducing spurious patterns. This demonstrates how neurosymbolic integration can convert complex predictors into transparent artefacts suitable for auditing, validation, and clinical decision support.

Use case (trustworthy medical imaging with neuro-symbolic pipelines). Brown et al. [BWAC+23] present the *SimpleMind* neuro-symbolic software framework for deploying trustworthy AI systems in real-world medical imaging.

- *Knowledge component.* Explicit rule-based models, structured clinical knowledge bases, and modular workflow definitions that encode radiological expertise and define permissible decision logic within imaging pipelines.

- *NeSy learning component.* Neural image analysis modules (*e.g.*, segmentation and classification networks) embedded within a symbolic orchestration layer that structures processing steps, constrains outputs, and enables human-interpretable inspection of intermediate decisions. In particular, here, the NeSy component shows useful when lack of annotated data is available to train a segmentation DNN but a KB on chest radiographs is available. Thanks to the latter, containing the protocol instructions on how to place the endotracheal tubes (to maintain airway patency and lung ventilation) and computing the area on which to apply DNN segmentation using spatial relationships in the semantic network, the NeSy approach assisted with alerts agreed by physicians.

- *Lessons learned.* Embedding neural image models within symbolic, modular pipelines increases transparency, debuggability, and regulatory readiness. The approach illustrates how neuro-symbolic design supports real-world deployment by aligning AI behavior with established clinical workflows and interpretability requirements.

3.3 Knowledge-grounded Clinical Dialogue Support

Clinical Dialogue Support refers to computing systems (*e.g.*, virtual doctors) that help clinicians or patients by having medically informed conversations, providing answers, guidance, or explanations grounded in reliable medical knowledge.

Use case (knowledge-grounded medical dialogue via heterogeneous graph reasoning). Liu et al. [LTLC21] propose a heterogeneous graph reasoning framework for medical dialogue systems that grounds conversational responses in structured medical knowledge.

- *Knowledge component.* A heterogeneous medical KG integrating diseases, symptoms, treatments, and dialogue entities to provide structured context for conversational reasoning.
- *NeSy learning component.* A graph reasoning module -fusing medical knowledge with entity context by attentional graph propagation performs multi-hop reasoning over the heterogeneous graph. Outputs are fused with neural dialogue generation models to produce knowledge-grounded responses.
- *Lessons learned.* Incorporating structured graph reasoning improves factual consistency and reduces hallucinations in medical dialogue generation. The work proposes a Medical Dialogue Consultant and Diagnosis Benchmark, and highlights the importance of symbolic grounding in safety-critical conversational AI.

Use case (Virtual doctor with dialogue knowledge-grounded augmented graphs). Varshney et al. [VZBE23] propose a dialogue-based, graph-based approach for knowledge-grounded medical dialogue generation.

- *Knowledge component.* Medical knowledge graphs (KGs) are enriched with contextual dialogue information and external medical resources (UMLS KBs) to support response generation.
- *NeSy learning component.* Graph neural encoders integrate structured medical knowledge with dialogue context representations, this time without attention mechanisms, guiding a neural language model to generate responses aligned with domain knowledge. In particular, the effective conversation understanding and generation is achieved through: 1) a masked entity dialogue model is finetuned on the dialogues to encode the augmented KG alongside the patient's utterances using a pretrained LLM; 2) The UMLS database is augmented with re-curated KG information from a medical entity prediction model to improve the semi-supervised medical entity annotation.
- *Lessons learned.* Augmenting neural dialogue models with structured KGs improves response relevance and medical correctness. The study reinforces that explicit symbolic grounding mitigates factual errors and enhances reliability in medical conversational systems.

3.4 Clinical Prediction & Monitoring

Clinical Prediction & Monitoring refers to the use of computing tools to track patient health and predict future medical events

Use case (neuro-symbolic patient monitoring). Fenske et al. [FBK23] propose a neuro-symbolic architecture for monitoring patients by integrating data-driven perception with formal symbolic reasoning to support continuous clinical assessment. The architecture testing a proof-of-concept example shows:

- *Knowledge component.* A symbolic knowledge base encoding medical monitoring rules, temporal constraints, and clinically relevant state transitions that define when patient conditions require attention or escalation.

- *NeSy learning component.* Neural components process heterogeneous patient data streams (*e.g.*, vital signs, sensor signals) to detect patterns and infer latent clinical states, which are then evaluated against symbolic rules through logical reasoning to derive high-level monitoring decisions.

- *Lessons learned.* The integration of symbolic medical rules with neural perception enables transparent, rule-grounded monitoring decisions while preserving adaptability to noisy real-world data. The work highlights how neuro-symbolic systems can ensure safety and explainability in continuous patient monitoring scenarios.

Use case (human-in-the-loop federated GNNs for disease classification). Hausleitner et al. [HMHP24] introduce a collaborative weighting mechanism for federated Graph Neural Networks in disease classification, incorporating expert feedback into distributed training.

- *Knowledge component.* Expert-informed weighting schemes and domain-relevant graph structures encoding clinically meaningful relationships among patient features.

- *NeSy learning component.* Federated GNN training across decentralized datasets, augmented with a human-in-the-loop mechanism that adjusts aggregation weights based on expert evaluation of intermediate representations and predictions.

- *Lessons learned.* Integrating expert-guided symbolic feedback into federated neural learning improves model robustness and trustworthiness without compromising data privacy. The work illustrates how human-guided symbolic interventions can stabilize distributed medical AI systems.

Use case (graph-based prediction of prognosis indicators in oncology). Bontempo et al. [BBL+23] propose GDS-MIL, a graph-based dual-stream multiple instance learning (MIL) framework to enhance the prediction of the prognosis indicator PFI (platinum-free interval) from whole-slide images (WSI) histopathology data.

- *Knowledge component.* Graph-structured representations of whole-slide images capturing spatial and morphological relationships between tissue regions. The graph abstraction encodes structural dependencies among cellular and tissue-level components relevant to tumor progression.

- *NeSy learning component.* A dual-stream MIL architecture processes both instance-level features (patch embeddings extracted from histopathology images) and graph-structured relational information. The prediction averages scores of instances and bag classifiers. The GNN's graph attention gate models inter-instance dependencies, enabling relational reasoning over tissue topology to inform PFI prediction.

- *Lessons learned.* Incorporating graph-structured relational inductive bias into MIL improves predictive performance compared to purely instance-based approaches. The study in particular demonstrates that explicitly modeling spatial and structural relationships between tissue regions enhances robustness and clinical relevance of oncological outcome prediction.

3.5 Computational pharmacology

Computational pharmacology refers to the use of computer models and artificial intelligence to study how drugs work in the body with a special focus on the drug's mechanism-of-action (MoA), that is, the molecular processes that lead to its medicinal effect. It helps scientists discover new drugs, predict how different medicines might interact, and identify possible side effects before testing in people.

Use case (drug discovery). MARS [DGG+25] represents a NeSy system that learns weighted logical rules (resembling MoAs) to explain drug discovery predictions.

- *Knowledge component.* A tailored MoA-net knowledge graphs, and the learned weighted rules.
- *NeSy learning component.* A deep reinforcement learning (RL) approach to train an agent that perform walks through the knowledge graph MoA-net. The graph traversal aim to link pairs of nodes that share a predefined target relation and will lead to the rule-learning.
- *Lessons learned.* The learned rules provide explainability for the predictions. MARS is short-cut aware to lead to rules that are more meaningful from a biomedical point of view.

Use case (drug repurposing hypothesis). Nunes et al. [NBP25] developed a method for generating scientific explanations of drug repurposing hypotheses in knowledge graphs. It employs reward and policy mechanisms that consider desirable properties of scientific explanation to guide a reinforcement learning agent in identifying explanatory paths.

- *Knowledge component.* Biomedical knowledge graphs such as Hetionet, PrimeKG, and OREGANO that capture information about drugs, diseases, and other entities relevant to drug repurposing. Enrichment with established ontologies including the National Cancer Institute Thesaurus (NCIT) and Chemical Entities of Biological Interest (ChEBI).
- *NeSy learning component.* A reinforcement learning approach guided by a dual reward that values both fidelity to the hypothesis and relevance to generate explanatory paths over the KGs. Relevance of paths is computed by computing the information content (IC) of their edges and nodes. OWL2Vec* embeddings are used [CHJ+21] to support the calculation of the IC of a node.
- *Lessons learned.* The method outperforms the state of the art and yields more relevant explanatory paths, which experts judge to be of higher quality.

Use case (chemical toxicity prediction). Glauer et al. [GNMH23] developed an approach to use the ontology hierarchy of a chemical ontology as symbolic knowledge to improve the generalisability of a subsequently trained neural network for the difficult task of toxicity prediction for unseen molecules.

- *Knowledge component.* The structure-based classification of chemicals in the ChEBI ontology [DHME09].
- *NeSy learning component.* A masked-language model pre-trained with ontology-based tasks then further trained for toxicity prediction tasks.
- *Lessons learned.* The model learns to focus attention on more meaningful chemical groups when making predictions with ontology pre-training than without, paving a path towards greater robustness and interpretability. In addition, the training time is reduced after ontology pre-training, indicating that the model is better placed to learn what matters for toxicity prediction with the ontology pre-training than without. This strategy has general applicability as a neuro-symbolic approach to embed meaningful semantics into neural networks

3.6 Clinical information extraction

Clinical information extraction is the process of automatically identifying and structuring important medical information from healthcare documents, such as doctors' notes, discharge summaries, or pathology reports.

Use case (neuro-symbolic named entity recognition in oncology notes). García-Barragán et al. [GBSV+25] introduce a Spanish-language NeuroSymbolic System for Cancer (NSSC), in order to enhance named entity recognition (NER) and entity linking in oncological clinical narratives.

- *Knowledge component.* Biomedical ontologies and controlled vocabularies used to validate, constrain, and disambiguate recognized entities, ensuring consistency with domain semantics and hierarchical concept structures.
- *NeSy learning component.* A neural entity recognition (NER) model extracts candidate entities (via Chain-of-Thought LLM prompting) from unstructured oncology notes, whose outputs are refined through symbolic reasoning to enforce semantic coherence and resolve ambiguities. Ontology-based linking mechanisms enhance interoperability and semantic alignment in clinical data, facilitating better integration with biomedical vocabularies like UMLS.
- *Lessons learned.* Combining large language models with ontology-guided post-processing significantly improves entity recognition accuracy and linking reliability compared to purely neural baselines. The system demonstrates that symbolic constraints are particularly valuable in high-stakes domains like oncology, where terminological precision and semantic consistency are critical.

3.7 Biomedical knowledge discovery

Biomedical knowledge discovery refers to the use of computer methods to find new insights and relationships in large collections of biological and medical data.

Use case (conceptual validation of GNNs in biomedical graphs). Finzel et al. [FSA+22] investigate how to generate symbolic explanations for Graph Neural Networks (GNNs) by extracting human-interpretable predicates from relevance-ranked subgraphs in biomedical settings.

- *Knowledge component.* Symbolic predicates derived from structured biomedical graphs, representing meaningful relational patterns among entities (*e.g.*, disease–gene–pathway relationships).
- *NeSy learning component.* A GNN identifies relevance-ranked subgraphs for predictions, after which symbolic rules are extracted to approximate learned relational patterns. The extracted predicates serve as interpretable surrogates of the GNN's internal reasoning.
- *Lessons learned.* Symbolic predicate extraction enables conceptual validation of GNN predictions by domain experts, bridging the gap between distributed graph representations and human-understandable reasoning. The work highlights the importance of post-hoc symbolic abstraction for validating graph-based predictors in biomedical domains.

Use case (neuro-symbolic genomic variant prioritization). Althagafi et al. [AZCH24] develop a neuro-symbolic framework for prioritizing genomic variants by integrating biological knowledge graphs with neural prediction models.

- *Knowledge component.* Biomedical ontologies and knowledge graphs encoding gene–disease–phenotype relationships, providing structured biological context for candidate variants.
- *NeSy learning component.* Neural models learn variant representations enriched with graph-derived embeddings and logical constraints, enabling knowledge-enhanced prioritization of clinically relevant variants. Concretely, the *EmbedPVP* approach [AZCH24] fuses different ontologies (gene, phenotype and anatomy ontologies) through knowledge-graph and ontology embeddings to prioritize gene variants.
- *Lessons learned.* Knowledge-grounded learning improves prioritization accuracy and enhances biological interpretability of predictions. The study demonstrates the value of embedding structured genotype–phenotype knowledge into neural models for precision medicine applications. This work goes beyond semantic similarity metrics for detecting genotype-phenotype relations to predict gene-diseases, because it introduces the ability to learn this similarity within a latent space that allows to model complex relationships, improve predictive performance, as well as to find candidate disease genes with no known or noisy phenotype.

3.8 Medical reporting

Medical report generation is the process of automatically creating clear, structured reports from medical data, such as imaging scans, lab results, or patient records.

Use case (neuro-symbolic spinal medical report generation). Han et al. [HWX+21] propose a unified framework combining neural learning and symbolic reasoning for automated spinal radiology report generation.

- *Knowledge component.* A structured medical knowledge base encoding spinal anatomical entities, pathological findings, and logical dependencies between abnormalities.
- *NeSy learning component.* A convolutional neural network extracts visual features from spinal images, which are aligned with symbolic abnormality representations. A reasoning module enforces logical consistency between detected findings before generating the final report. The neural learning and symbolic reasoning system for spinal medical report generation are unified through an adversarial graph network that leverages a symbolic graph reasoner to perform complex semantic segmentation of spinal structures. Then a human-like symbolic module analyses causal effects of abnormal detected entities via meta-interpretive learning to propose relationship (causal effect) hypotheses.
- *Lessons learned.* Integrating symbolic reasoning into report generation improves logical consistency and reduces contradictory findings compared to purely neural generation models. The study demonstrates that neuro-symbolic constraints are essential for ensuring clinically coherent diagnostic narratives.

4 Discussion and Future Outlook

In this white paper, we have introduced neurosymbolic approaches in medicine and surveyed some recent examples of applications across medical evidence, clinical decision making, pharmacology, clinical prediction, clinical dialogue support, clinical information extraction, biomedical knowledge discovery, and medical reporting. Large language models

and multimodal foundation models are increasingly being adopted throughout medical and clinical informatics applications. Neurosymbolic approaches complement and extend the approaches of foundation models, addressing specific gaps that foundation models are not elegantly able to address, such as reliability, transparent explainability and reproducibility.

Some areas that require attention in future work are architectural complexity and the associated need for tooling, and the challenge of incomplete knowledge. These are elaborated below.

4.1 Architectural Complexity and the Need for Accessible Tooling

Neurosymbolic systems, by their nature, are more complex than either purely neural or purely symbolic approaches. Integrating learned representations with structured reasoning requires careful decisions at the interface between subsystems: how symbolic knowledge is encoded and updated, how neural outputs are mapped to symbolic predicates, how uncertainty propagates across the boundary, and how training signals flow back through components that may not be differentiable. This architectural heterogeneity imposes a significant engineering burden on practitioners. A clinician or biomedical researcher with domain expertise but limited machine learning background cannot reasonably be expected to construct, tune, or maintain such a pipeline without substantial technical support.

This complexity represents a structural barrier to adoption that is particularly acute in medicine, where the development and validation of decision-support tools must involve clinical domain experts alongside engineers. Current implementations of NeSy systems tend to be bespoke, tightly coupled to specific knowledge representations and reasoning frameworks, and difficult to transfer across institutions or clinical contexts. The absence of mature, well-documented, and modular tooling analogous to what exists in the deep learning ecosystem means that each deployment requires substantial re-engineering effort.

Addressing this gap is a prerequisite for broad clinical uptake. Promising steps include the development of declarative interfaces that allow domain experts to specify symbolic constraints and knowledge without writing inference code, as well as modular architectures that allow individual components — the knowledge graph, the neural encoder, the reasoning layer — to be swapped or updated independently. Standardized benchmarks specifically targeting neurosymbolic medical AI would further accelerate progress by enabling systematic comparison of design choices. Without deliberate investment in tooling and abstraction layers, neurosymbolic approaches risk remaining a research curiosity rather than a clinical utility.

4.2 Knowledge Incompleteness and Explainability

A central promise of neurosymbolic approaches in medicine is the ability to generate human-interpretable explanations grounded in established biomedical knowledge. When a model flags a drug interaction or suggests a diagnosis, the reasoning chain can in principle be audited, challenged, and communicated to a clinician. This transparency is one of the most compelling arguments for neurosymbolic architectures over opaque neural alternatives.

However, biomedical knowledge graphs are necessarily incomplete, which may hinder their use in explanations in some cases. The landscape of disease mechanisms, gene-phenotype associations, drug targets, and clinical evidence is vast, rapidly evolving, and unevenly covered. Edges that do not yet exist in a knowledge graph cannot support an explanatory reasoning chain. This limitation is not unique to neurosymbolic approaches: any knowledge-based system will inherit the biases and gaps of its knowledge sources. There is work that makes use of neurosymbolic approaches for extracting relations between entities from text. These entities can be within a sentence [JSM23], across paragraphs [JMKS24], or documents [JMSK24]. Relation extraction techniques can be used for knowledge graph completion by considering the entities as the nodes and the relations as the edges. Al-

though this is possible, challenges exist in the accuracy, generalizability, and scalability of these approaches.

In the future, ideally, systems must be designed to communicate the knowledge status of their explanations, distinguishing between conclusions supported by well-established knowledge, those derived from weak or indirect evidence, and those for which no supporting knowledge currently exists.

Practical mitigations include integrating uncertainty quantification into the reasoning layer, flagging predictions that rely on low-confidence or sparse knowledge regions, and linking explanations to the evidence underlying each knowledge graph edge (*e.g.*, published evidence for the relationship).

Longer term, active learning and knowledge graph completion methods may allow the system to identify and prioritise its own knowledge gaps, directing curation effort toward the regions most consequential for clinical decision-making. In this respect, LLMs may themselves be purposed as aides to knowledge extraction that feed into symbolic frameworks, leading to positive improvement feedback cycles between learning, knowledge and reasoning.

Acknowledgements

This work has been supported by The Academy of Medical Sciences Networking Grant, Neurosymbolic AI for Medicine, NGR1\1857: https://neuro-symbolic-ai-medicine.github.io/.

References

[AAAA+23] Alaa Abd-Alrazaq, Rawan AlSaad, Dari Alhuwail, Arfan Ahmed, Padraig Mark Healy, Syed Latifi, Sarah Aziz, Rafat Damseh, Sadam Alabed Alrazak, and Javaid Sheikh. Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. *JMIR medical education*, 9:e48291, June 2023.

[AZCH24] Azza Althagafi, Fernando Zhapa-Camacho, and Robert Hoehndorf. Prioritizing genomic variants through neuro-symbolic, knowledge-enhanced learning. *Bioinformatics*, 40(5):btae301, 2024.

[BBL+23] Gianpaolo Bontempo, Nicola Bartolini, Marta Lovino, Federico Bolelli, Anni Virtanen, and Elisa Ficarra. Enhancing pfi prediction with gds-mil: A graph-based dual stream mil approach. In *International Conference on Image Analysis and Processing*, pages 550–562. Springer, 2023.

[BH05] Sebastian Bader and Pascal Hitzler. Dimensions of neural-symbolic integration - A structured survey. In Sergei N. Artëmov, Howard Barringer, Artur S. d'Avila Garcez, Luís C. Lamb, and John Woods, editors, *We Will Show Them! Essays in Honour of Dov Gabbay, Volume One*, pages 167–194. College Publications, 2005.

[BP24] Jean-Christophe Bélisle-Pipon. Why we need to be careful with LLMs in medicine. *Frontiers in Medicine*, 11, December 2024. Publisher: Frontiers.

[BPP+26] Andrew M. Bean, Rebecca Elizabeth Payne, Guy Parsons, Hannah Rose Kirk, Juan Ciro, Rafael Mosquera-Gómez, Sara Hincapié M, Aruna S. Ekanayaka, Lionel Tarassenko, Luc Rocher, and Adam Mahdi. Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. *Nature Medicine*, pages 1–7, February 2026.

[BWAC+23] Matthew S Brown, M Wasil Wahi-Anwar, Youngwon Choi, Morgan Daly, Liza Shrestha, Koon-Pong Wong, Jonathan G Goldin, and Dieter R Enzmann. Implementing trustworthy ai in real-world medical imaging using the simplemind software environment. In *NeSy*, pages 195–203, 2023.

[BWS+23] Manuela Benary, Xing David Wang, Max Schmidt, Dominik Soll, Georg Hilfenhaus, Mani Nassir, Christian Sigler, Maren Knödler, Ulrich Keller, Dieter Beule, Ulrich Keilholz, Ulf Leser, and Damian T. Rieke. Leveraging Large Language Models for Decision Support in Personalized Oncology. *JAMA Network Open*, 6(11):e2343689, November 2023.

[CD22] Andrew Cropper and Sebastijan Dumancic. Inductive logic programming at 30: A new introduction. *J. Artif. Intell. Res.*, 74:765–850, 2022.

[CDH+23] Jiaoyan Chen, Hang Dong, Janna Hastings, Ernesto Jiménez-Ruiz, Vanessa López, Pierre Monnin, Catia Pesquita, Petr Skoda, and Valentina Tamma. Knowledge graphs for the life sciences: Recent developments, challenges and opportunities. *TGDK*, 1(1):5:1–5:33, 2023.

[CHJ+21] Jiaoyan Chen, Pan Hu, Ernesto Jiménez-Ruiz, Ole Magnus Holter, Denvar Antonyrajah, and Ian Horrocks. Owl2vec*: embedding of OWL ontologies. *Mach. Learn.*, 110(7):1813–1845, 2021.

[CKM+23] Jan Clusmann, Fiona R. Kolbinger, Hannah Sophie Muti, Zunamys I. Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P. Veldhuizen, Sophia J. Wagner, and Jakob Nikolas Kather. The future landscape of large language models in medicine. *Communications Medicine*, 3(1):1–8, October 2023. Number: 1 Publisher: Nature Publishing Group.

[CWBdPM20] Roberto Confalonieri, Tillman Weyde, Tarek R. Besold, and Fermín Moscoso del Prado Martín. TREPAN reloaded: A knowledge-driven approach to explaining black-box models. In Giuseppe De Giacomo, Alejandro Catalá, Bistra Dilkina, Michela Milano, Senén Barro, Alberto Bugarín, and Jérôme Lang, editors, *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2457–2464. IOS Press, 2020.

[CWBM21] Roberto Confalonieri, Tillman Weyde, Tarek R. Besold, and Fermín Moscoso del Prado Martín. Using ontologies to enhance human understandability of global post-hoc explanations of black-box models. *Artificial Intelligence*, 296:103471, 2021.

[CWGS+20] Denise Calaprice-Whitty, Karim Galil, Wael Salloum, Ashkon Zariv, and Bernal Jimenez. Improving clinical trial participant prescreening with artificial intelligence (ai): a comparison of the results of ai-assisted vs standard methods in 3 oncology trials. *Therapeutic innovation & regulatory science*, 54:69–74, 2020.

[DGG+25] Lauren Nicole DeLong, Yojana Gadiya, Paola Galdi, Jacques D. Fleuriot, and Daniel Domingo-Fernández. MARS: A neurosymbolic approach for interpretable drug discovery, 2025. arXiv:2410.05289 [cs] version: 3.

[dGL23] Artur d'Avila Garcez and Luís C. Lamb. Neurosymbolic AI: the 3rd wave. *Artif. Intell. Rev.*, 56(11):12387–12406, 2023.

[dGZ99] Artur S. d'Avila Garcez and Gerson Zaverucha. The connectionist inductive learning and logic programming system. *Appl. Intell.*, 11(1):59–77, 1999.

[DHME09] Kirill Degtyarenko, Janna Hastings, Paula de Matos, and Marcus Ennis. ChEBI: An Open Bioinformatics and Cheminformatics Resource. *Current Protocols in Bioinformatics*, 26(1):14.9.1–14.9.20, 2009. _eprint: https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/0471250953.bi1409s26.

[DHP+25] Fabio Dennstädt, Janna Hastings, Paul Martin Putora, Max Schmerder, and Nikola Cihoric. Implementing large language models in healthcare while balancing control, collaboration, costs and security. *npj Digital Medicine*, 8(1):1–4, March 2025.

[DHW+25] Fabio Dennstädt, Janna Hastings, Paul Windisch, Aleksa Jovanovic, Tijana Žunić Marić, Sarah Brüningk, Daniel M. Aebersold, Antje Knopf, and Nikola Cihoric. The EU AI Act: Implications and Compliance Guidance for Healthcare Facilities, November 2025.

[DZP+24] Fabio Dennstädt, Johannes Zink, Paul Martin Putora, Janna Hastings, and Nikola Cihoric. Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain. *Systematic Reviews*, 13(1):158, June 2024.

[EJG+23] Daniel E. Ehrmann, Shalmali Joshi, Sebastian D. Goodfellow, Mjaye L. Mazwi, and Danny Eytan. Making machine learning matter to clinicians: model actionability in medical decision-making. *npj Digital Medicine*, 6(1):1–5, January 2023. Number: 1 Publisher: Nature Publishing Group.

[FBK23] Ole Fenske, Sebastian Bader, and Thomas Kirste. Neuro-symbolic artificial intelligence for patient monitoring. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 19–25. Springer, 2023.

[FSA+22] Bettina Finzel, Anna Saranti, Alessa Angerschmid, David Tafler, Bastian Pfeifer, and Andreas Holzinger. Generating explanations for conceptual validation of graph neural networks: An investigation of symbolic predicates learned on relevance-ranked sub-graphs. *KI-Künstliche Intelligenz*, 36(3):271–285, 2022.

[GBSV+25] Álvaro García-Barragán, Ahmad Sakor, Maria-Esther Vidal, Ernestina Menasalvas, Juan Cristobal Sanchez Gonzalez, Mariano Provencio, and Víctor Robles. Nssc: a neuro-symbolic ai system for enhancing accuracy of named entity recognition and linking from oncologic clinical notes. *Medical & Biological Engineering & Computing*, 63(3):749–772, 2025.

[GLH+25] Ziyi Guan, Jason Chun Lok Li, Zhijian Hou, Pingping Zhang, Donglai Xu, Yuzhi Zhao, Mengyang Wu, Jinpeng Chen, Thanh-Toan Nguyen, Pengfei Xian, Wenao Ma, Shengchao Qin, Graziano Chesi, and Ngai Wong. KG-RAG: enhancing GUI agent decision-making via knowledge graph-driven retrieval-augmented generation. *CoRR*, abs/2509.00366, 2025.

[GNMH23] Martin Glauer, Fabian Neuhaus, Till Mossakowski, and Janna Hastings. Ontology Pre-training for Poison Prediction. In Dietmar Seipel and Alexander Steen, editors, *KI 2023: Advances in Artificial Intelligence*, pages 31–45, Cham, 2023. Springer Nature Switzerland.

[GWMH22] Martin Glauer, Robert West, Susan Michie, and Janna Hastings. ESC-Rules: Explainable, Semantically Constrained Rule Sets. In *Proceedings of the 16th International Workshop on Neural-Symbolic Learning and Reasoning as part of the 2nd International Joint Conference on Learning & Reasoning (IJCLR 2022), Cumberland Lodge, Windsor Great Park, UK, September 28-30, 2022*, volume 3212 of *CEUR Workshop Proceedings*, pages 94–103. CEUR-WS.org, 2022.

[Has24] Janna Hastings. Preventing harm from non-conscious bias in medical generative AI. *The Lancet Digital Health*, 6(1):e2–e3, January 2024.

[HDMN25] Pascal Hitzler, Abhilekha Dalal, Mohammad Saeid Mahdavinejad, and Sanaz Saki Norouzi, editors. *Handbook on Neurosymbolic AI and Knowledge Graphs*, volume 400 of *Frontiers in Artificial Intelligence and Applications*. IOS Press / Sage, 2025.

[Hit21] Pascal Hitzler. A review of the semantic web field. *Commun. ACM*, 64(2):76–83, 2021.

[HJRW25] David Herron, Ernesto Jiménez-Ruiz, and Tillman Weyde. On the Potential of Logic and Reasoning in Neurosymbolic Systems using OWL-based Knowledge Graphs. *Neurosymbolic Artificial Intelligence*, 1, 2025. https://journals.sagepub.com/doi/epdf/10.1177/29498732251320043.

[HMHP24] Christian Hausleitner, Heimo Mueller, Andreas Holzinger, and Bastian Pfeifer. Collaborative weighting in federated graph neural networks for disease classification with the human-in-the-loop. *Scientific Reports*, 14(1):21839, 2024.

[HSS+25] Pascal Hitzler, Cogan Matthew Shimizu, Daria Stepanova, Frank van Harmelen, et al. Seminar 25291 on (Actual) Neurosymbolic AI: Combining Deep Learning and Knowledge Graphs. *Dagstuhl Reports*, 2025.

[HWX+21] Zhongyi Han, Benzheng Wei, Xiaoming Xi, Bo Chen, Yilong Yin, and Shuo Li. Unifying neural learning and symbolic reasoning for spinal medical report generation. *Medical image analysis*, 67:101872, 2021.

[IHvH+25] Filip Ilievski, Barbara Hammer, Frank van Harmelen, Benjamin Paassen, Sascha Saralajew, Ute Schmid, Michael Biehl, Marianna Bolognesi, Xin Luna Dong, Kiril Gashteovski, Pascal Hitzler, Giuseppe Marra, Pasquale Minervini, Martin Mundt, Axel-Cyrille Ngonga Ngomo, Alessandro Oltramari, Gabriella Pasi, Zeynep G. Saribatur, Luciano Serafini, John Shawe-Taylor, Vered Shwartz, Gabriella Skitalinskaya, Clemens Stachl, Gido M. van de Ven, and Thomas Villmann. Aligning generalization between humans and machines. *Nature Machine Intelligence*, 7:1378–1389, 2025.

[JMKS24] Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru, and Kuldeep Singh. Revisiting document-level relation extraction with context-guided link prediction. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, February 20-27, 2024, Vancouver, Canada*, pages 18327–18335. AAAI Press, 2024.

[JMSK24] Monika Jain, Raghava Mutharaju, Kuldeep Singh, and Ramakanth Kavuluru. Knowledge-driven cross-document relation extraction. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and*

virtual meeting, August 11-16, 2024, volume ACL 2024 of *Findings of ACL*, pages 3787–3797. Association for Computational Linguistics, 2024.

[JSM23] Monika Jain, Kuldeep Singh, and Raghava Mutharaju. Reonto: A neuro-symbolic approach for biomedical relation extraction. In Danai Koutra, Claudia Plant, Manuel Gomez Rodriguez, Elena Baralis, and Francesco Bonchi, editors, *Machine Learning and Knowledge Discovery in Databases: Research Track - European Conference, ECML PKDD 2023, Turin, Italy, September 18-22, 2023, Proceedings, Part IV*, volume 14172 of *Lecture Notes in Computer Science*, pages 230–247. Springer, 2023.

[KTK+21] Tian Kang, Ali Turfah, Jaehyun Kim, Adler Perotte, and Chunhua Weng. A neuro-symbolic method for understanding free-text medical evidence. *Journal of the American Medical Informatics Association: JAMIA*, 28(8):1703–1711, July 2021.

[LBH10] Jens Lehmann, Sebastian Bader, and Pascal Hitzler. Extracting reduced logic programs from artificial neural networks. *Appl. Intell.*, 32(3):249–266, 2010.

[LG24] Ben Li and Stephen Gilbert. Artificial Intelligence awarded two Nobel Prizes for innovations that will shape the future of medicine. *npj Digital Medicine*, 7(1):1–3, November 2024. Publisher: Nature Publishing Group.

[LTLC21] Wenge Liu, Jianheng Tang, Xiaodan Liang, and Qingling Cai. Heterogeneous graph reasoning for knowledge-grounded medical dialogue system. *Neurocomputing*, 442:260–268, 2021.

[LYK+20] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240, 2020.

[MWF+20] Susan Michie, Robert West, Ailbhe N. Finnerty, Emma Norris, Alison J. Wright, Marta M. Marques, Marie Johnston, Michael P. Kelly, James Thomas, and Janna Hastings. Representation of behaviour change interventions and their evaluation: Development of the Upper Level of the Behaviour Change Intervention Ontology. *Wellcome Open Research*, 5:123, June 2020.

[NBP25] Susana Nunes, Samy Badreddine, and Catia Pesquita. Rewarding explainability in drug repurposing with knowledge graphs. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, pages 4624–4632. ijcai.org, 2025.

[NGT22] Kwun Ho Ngan, Artur D'avila Garcez, and Joseph Townsend. Extracting meaningful High-Fidelity knowledge from convolutional neural networks. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–17, July 2022.

[NMBP+24] Kwun Ho Ngan, Esma Mansouri-Benssassi, James Phelan, Joseph Townsend, and Artur d'Avila Garcez. From explanation to intervention: Interactive knowledge extraction from convolutional neural networks used in radiology. *PLOS ONE*, 19(4):1–29, 04 2024.

[PF25] Sandro Preto and Marcelo Finger. Effective reasoning over neural networks using Łukasiewicz logic. In Pascal Hitzler, Abhilekha Dalal, Mohammad Saeid Mahdavinejad, and Sanaz Saki Norouzi, editors, *Handbook on Neurosymbolic AI and Knowledge Graphs*, volume 400 of *Frontiers in Artificial Intelligence and Applications*, pages 609–630. IOS Press / Sage, 2025.

[PPP20] Cecilia Panigutti, Alan Perotti, and Dino Pedreschi. Doctor XAI: An ontology-based approach to black-box sequential data classification explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*, pages 629–639. ACM, 2020.

[SZEH21] Md. Kamruzzaman Sarker, Lu Zhou, Aaron Eberhart, and Pascal Hitzler. Neuro-symbolic artificial intelligence: Current trends. *AI Commun.*, 34(3):197–209, 2021.

[TKI21] Joe Townsend, Theodoros Kasioumis, and Hiroya Inakoshi. ERIC: Extracting relations inferred from convolutions. In *Computer Vision – ACCV 2020*, Lecture notes in computer science, pages 206–222. Springer International Publishing, Cham, 2021.

[VAM24] M. Vaccaro, A. Almaatouq, and T. Malone. When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nat Hum Behav*, 8:2293–2303, 2024.

[vHtT19] Frank van Harmelen and Annette ten Teije. A boxology of design patterns forhybrid learningand reasoning systems. *J. Web Eng.*, 18(1-3):97–124, 2019.

[VZBE23] Deeksha Varshney, Aizan Zafar, Niranshu Kumar Behera, and Asif Ekbal. Knowledge grounded medical dialogue generation using augmented graphs. *Scientific Reports*, 13(1):3310, 2023.

[WDWK19] Po-Wei Wang, Priya L. Donti, Bryan Wilder, and J. Zico Kolter. Satnet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6545–6554. PMLR, 2019.

[WZV+23] Felix Wong, Erica J. Zheng, Jacqueline A. Valeri, Nina M. Donghia, Melis N. Anahtar, Satotaka Omori, Alicia Li, Andres Cubillos-Ruiz, Aarti Krishnan, Wengong Jin, Abigail L. Manson, Jens Friedrichs, Ralf Helbig, Behnoush Hajian, Dawid K. Fiejtek, Florence F. Wagner, Holly H. Soutter, Ashlee M. Earl, Jonathan M. Stokes, Lars D. Renner, and James J. Collins. Discovery of a structural class of antibiotics with explainable deep learning. *Nature*, pages 1–9, December 2023. Publisher: Nature Publishing Group.