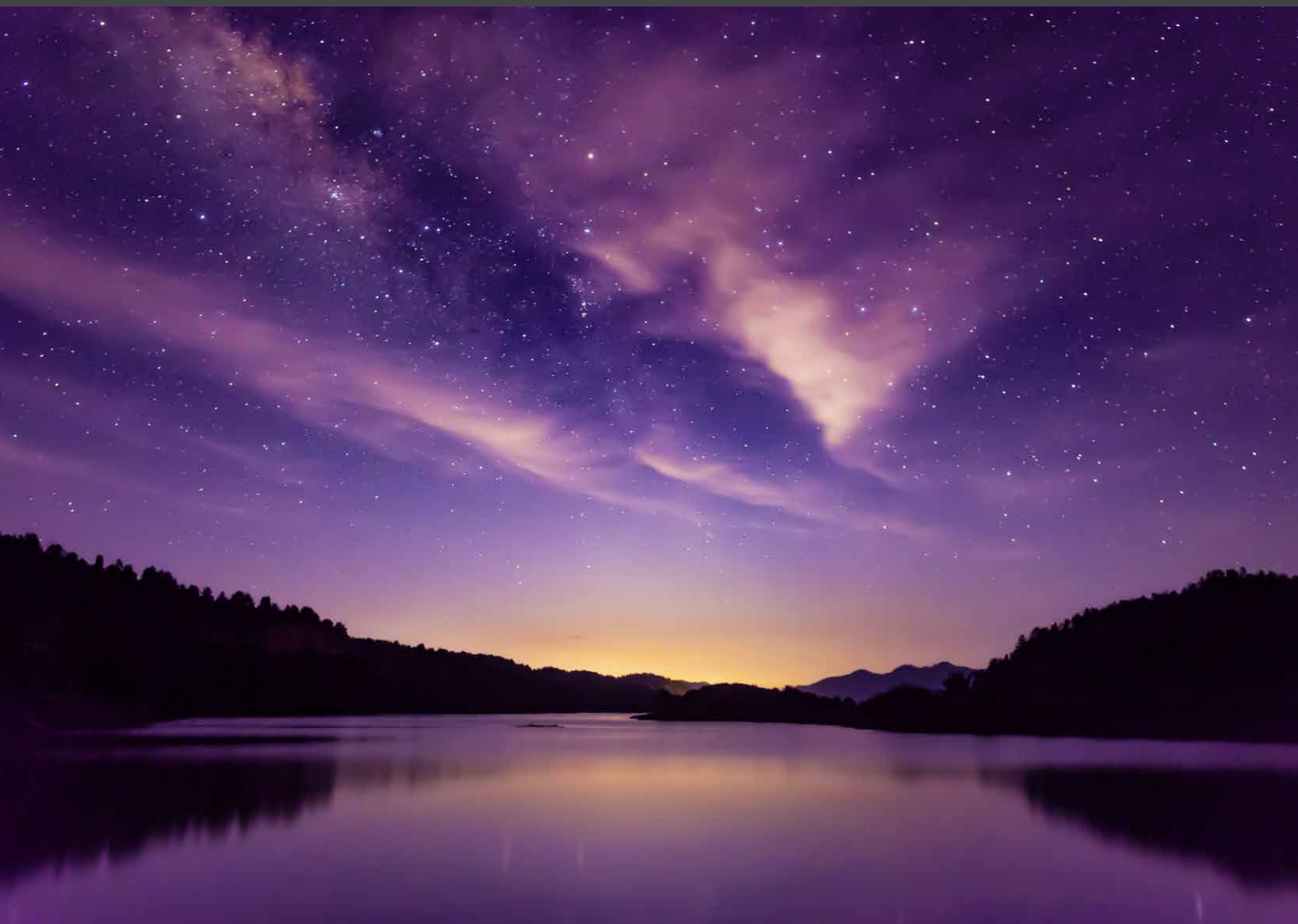


Use case

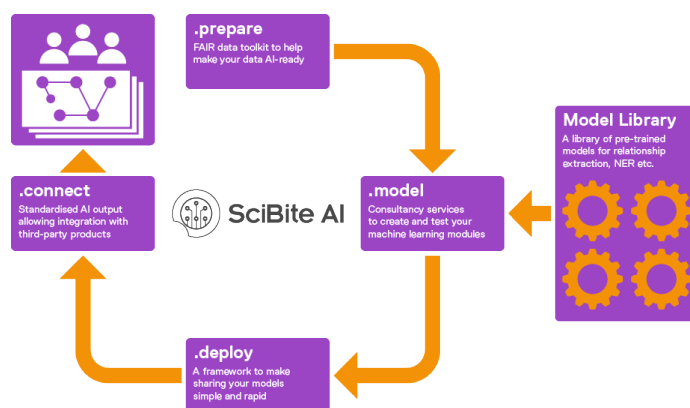
Leveraging Machine Learning Models: Combining Deep Learning with Semantics Technologies



Leveraging Machine Learning Models: Combining Deep Learning with Semantics Technologies

Deep learning models can be trained to understand language and the context of how words are used in sentences. Instead of providing an algorithm with rules, they can be taught using examples which they can subsequently generalise and learn from. It is this emphasis on pattern recognition that enables them to be applied to situations where pre-defined rules don't exist. However, the accuracy of such models is highly dependent on the quality of the training data used to build them.

SciBite AI combines the flexibility of deep learning pattern recognition with the reliability of SciBite's semantic technologies. The use cases below highlight the power of SciBite AI, which provides a framework to incorporate different Machine Learning approaches, ensuring that it can be applied to a wide range of problems.

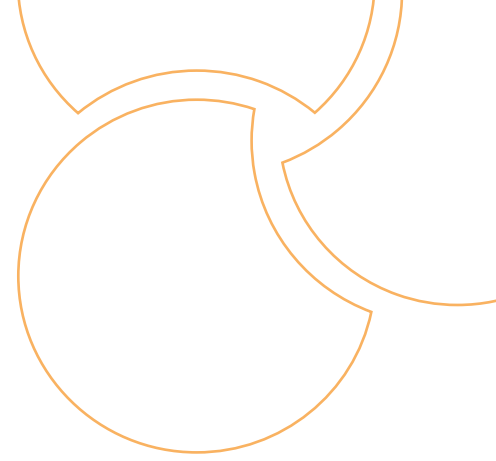


Model Type: Named Entity Recognition (NER)

SciBite's ultra-fast Named Entity Recognition (NER) engine, [TERMite](#), uses a rules-based approach to identify and extract scientific metadata from unstructured text. Rule-based approaches are not only repeatable, they also enable users to easily understand why something has been found. For TERMite, these rules are defined in the form of hand-curated ontologies which offer the advantage of defining the semantics of the entities being identified so that result sharing is meaningful. There are also challenges in using these approaches. Firstly, describing all entities and all their possible synonyms for NER requires considerable manual curation – something SciBite has invested much expertise into creating for users. Secondly, they can only identify known entities contained in these ontologies or rules. Although it is possible to create new rules, it is not always practical to create rules that cover everything.

We used SciBite AI to train the deep learning algorithm BioBERT with a set of articles that had been annotated with TERMite using SciBite's ontologies. This offers two key advantages. Firstly, semantic enrichment of the training data enables the models to understand the scientific language in these articles and the context in which it is used, based on which they can generalise. Secondly, we embed some of our vast ontology understanding into these models through the training process which helps the model to learn from the rules used in TERMite.

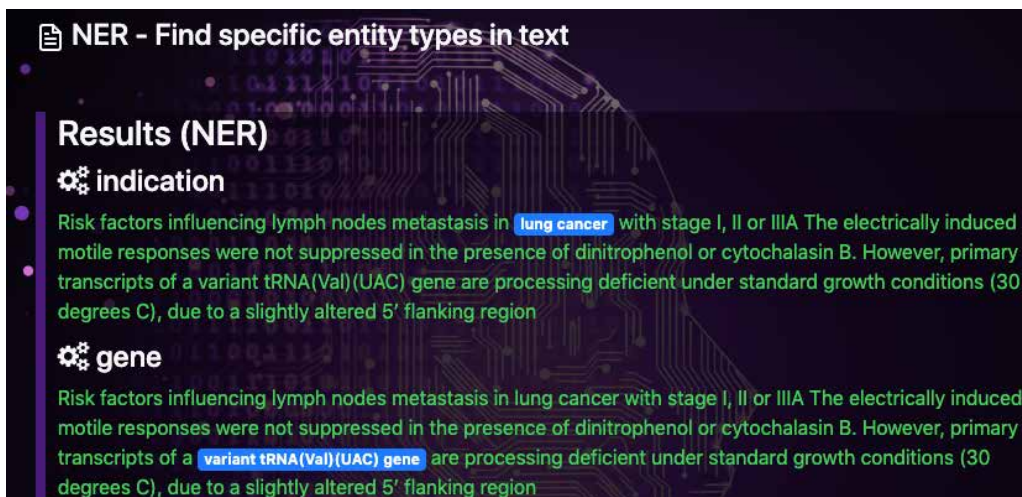
We then compared the results of applying i) BioBERT, ii) TERMite and iii) SciBite AI to the following example sentences with the goal of identifying the gene entities, which are highlighted in blue.



Example 1: 'Effects of a selective **monoamine oxidase (MAO)--A inhibitor**, clorgyline, a selective **MAO-B** inhibitor, deprenyl, and a non-selective **MAO** inhibitor, nialamide, were investigated on footshock-induced aggression (FIA) in paired rats.'

Method Used	Entities Correctly Identified as Genes	Entities Incorrectly Identified as Genes	Other Errors
BioBERT	1. MAO-B 2. MAO	monoamine	oxidase (MAO)—A Identified as a drug
TERMite	1. monoamine oxidase (MAO)--A 2. MAO-B	None	None
SciBite AI	1. monoamine oxidase (MAO)--A 2. MAO-B 3. MAO	None	None

Example 2: 'Risk factors influencing lymph nodes metastasis in lung cancer with stage I, II or IIIA The electrically induced motile responses were not suppressed in the presence of dinitrophenol or cytochalasin B. However, primary transcripts of a **variant tRNA(Val) (UAC) gene** are processing deficient under standard growth conditions (30 degrees C), due to a slightly altered 5' flanking region.'



NER - Find specific entity types in text

Results (NER)

indication
Risk factors influencing lymph nodes metastasis in **lung cancer** with stage I, II or IIIA The electrically induced motile responses were not suppressed in the presence of dinitrophenol or cytochalasin B. However, primary transcripts of a variant tRNA(Val)(UAC) gene are processing deficient under standard growth conditions (30 degrees C), due to a slightly altered 5' flanking region

gene
Risk factors influencing lymph nodes metastasis in lung cancer with stage I, II or IIIA The electrically induced motile responses were not suppressed in the presence of dinitrophenol or cytochalasin B. However, primary transcripts of a **variant tRNA(Val)(UAC) gene** are processing deficient under standard growth conditions (30 degrees C), due to a slightly altered 5' flanking region

Method Used	Entities Correctly Identified as Genes	Entities Incorrectly Identified as Genes	Other Errors
BioBERT	1. variant tRNA(Val)(UAC) gene	1. cytochalasin B	None
TERMite	None	None	None
SciBite AI	1. variant tRNA(Val)(UAC) gene	None	None

Example 3: 'Such mutations are thought to exert their dominant phenotype by sequestration of the **guanine nucleotide exchange factor (GNEF)**.'

Method Used	Entities Correctly Identified as Genes	Entities Incorrectly Identified as Genes	Other Errors
BioBERT	1. GNEF	1. guanine nucleotide	Exchange factor was incorrectly identified as a Drug/Chemical
TERMite	None	None	None
SciBite AI	1. guanine nucleotide exchange factor 2. GNEF	None	None

These examples clearly illustrate that, while deep learning models such as BioBERT can correctly identify scientific terms within unstructured text, including those that are not recognised by TERMite, they can struggle to correctly assign that term to a specific type of entity and so issues with false positives and incorrect classification persist. Conversely, while TERMite did not identify all possible entities, those that were found were accurately assigned to a specific gene.

However, by combining the accuracy of a rules based system with a deep learning model that has been designed to understand language, SciBite AI was able to accurately identify all true positives in the example sentences without any errors.

Through the unique and powerful combination of semantic technologies and deep learning, SciBite AI can help ensure all references to entities of interest and their synonyms are identified accurately, regardless of whether they are present in an ontology or not.

By identifying potential new entities and synonyms, SciBite AI also helps SciBite augment and extend our existing ontologies and rapidly develop ontologies for new domains, which can in turn be used to further train SciBite AI.

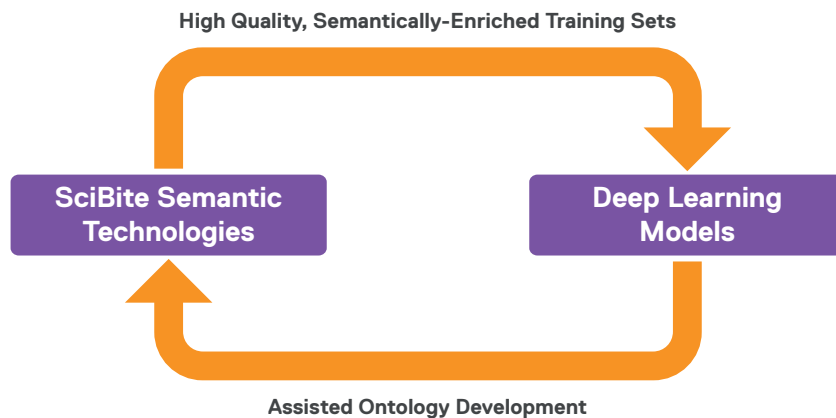


Figure: Combining the power of SciBite's semantic technologies with cutting-edge AI techniques

For example, new terms identified via SciBite AI can be easily added to an existing ontology managed by [CENtree](#). The updated ontology can be pushed to TERMite with a single click, enabling TERMite to evolve and identify the new terms in all future text analyses.

Use Case: Competitive Intelligence

Monitoring the publication and patent activity of others is essential for several common business processes:

- Maintaining awareness of technical developments in a particular technical field
- Identifying potential partners or in-licensing opportunities
- Early identification of potential patent infringements

This involves scanning through Pubmed, publicly available grant information, patent databases and a range of biotech-focused news websites in an attempt to identify articles of interest amongst the background 'noise'.

Manually searching different sources with multiple keywords or phrases of interest is resource intensive,

resulting in constraints on the number of sources that can be scanned, the frequency at which the scanning can be performed and the depth of review possible for potentially interesting articles. Given the manual nature of this work, there is a strong possibility that something will be missed. The very nature of attempting to identify interesting innovations also presents a challenge. For example, when a novel drug or new brand name for a known drug is used in an article for the first time, then by its very nature it will not yet be present in an ontology and therefore will not be identified by a computational method that relies purely on a rules based approach.



SciBite's semantic technologies have already been successfully applied to novelty scanning and competitor analysis within the Pharmaceutical industry, where they have been demonstrated to significantly reduce the duration of the competitive scanning and review process, such as [LifeArc's Data Mining and Horizon Scanning initiative](#). SciBite AI takes this to the next level, enabling models to not only identify potentially novel uses of existing things, but also to identify articles or patents

where novel genes, drugs or technologies are mentioned for the first time.

Through the unique and powerful combination of semantic technologies and deep learning, SciBite AI can help Pharmaceutical companies streamline competitive scanning activities and identify references to existing entities of interest as well as those that are truly novel.

Model Type: Relationship Extraction

Biomedical texts often describe entities, such as genes, drugs and diseases, in relation to one another, such as causation and treatment. However in the majority of cases, the co-occurrence of two entities with a sentence is not indicative of a causal relationship. Conversely, where a relationship does exist it is often described using complex language, which can be challenging even for expert curators and impenetrable to many computational methods.

different ways of describing the same concept can also be aggregated or 'bundled' and run across the same data simultaneously. These patterns are not only simple to define, they also ensure TExpress delivers repeatable results which can be understood by whoever reviews them. However, continually adding new patterns to bundles in an attempt to cover every possible verb or edge case can have diminishing returns.

TExpress is an application of TERMite which employs a rules-based approach to extract relationships from unstructured text. In the case of TExpress, these rules are encoded in semantic patterns in the form Gene-Verb-Disease. Multiple semantic patterns which encompass

For example, when using a pattern to identify 2 genes (highlighted in green) and one bioverb (highlighted in blue) in the same sentence, TExpress recognises both of the following sentences.

Sentence 1

S6k1 regulates estrogen receptor alpha (eralpha) by phosphorylating it on serine 167, leading to transcriptional activation of eralpha.

Sentence 2

NF13 inhibits apoptosis without affecting Bcl-xL expression.

However, additional grammatical understanding is required to differentiate between positive sentences that describe a causal relationship (such as Sentence 1) and negative sentences that do not (such as Sentence 2). This is where deep learning models can help.

However, when applied to entire documents, using such models alone can take a long time. To overcome this, we applied TExpress to rapidly identify sentences that contain each of the elements of interest (such as combinations of two gene entities and a relevant bioverb). The trained model was then applied to leverage its grammatical capability and differentiate between positive and negative hits.

We used TERMite to identify candidate sentences to form the basis of the training data set and these were subsequently manually curated by experts in the field. This provided us with gold standard data, including both positive and negative sentences. This was used to train BioBERT to make generalisations about the way relationships are described in these sentences and also to learn based on the rich knowledge embedded in the ontologies. Hence, when it is exposed to a new sentence it is able to predict which category it belongs in.

This synergy between semantic technologies and deep learning is described in detail in our recent paper [Optimising biomedical relationship extraction with BioBERT](#).

Use Case: Adverse Event Detection

Adverse event reporting is a regulatory requirement in most countries. Pharmaceutical companies are obliged to scan a range of data sources for possible adverse events associated with their drugs, assess their validity and notify regulatory bodies based on severity.

Historically Pharmaceutical companies have relied on experts to read articles describing potential drug-adverse event relationships and assess their importance. But reading hundreds or even thousands of articles from start to finish to identify potential adverse events is clearly not a scalable solution. Reviewers simply don't have the time to read all of the available content relevant to the drug(s) of interest and the serious implications of missing a critical signal results in an incredibly risk averse process, generating many false positives.

The accurate identification of an actual adverse event is critically dependent on the context in which it is

Example 1: 'Intravenous fentanyl may produce **bradycardia**. Use Lazanda with caution in patients with bradyarrhythmias.'

TERMite's rules-based approach accurately identifies 'bradycardia' and 'bradyarrhythmias' as terms that are matched within SciBite's adverse event dictionary. However, the second sentence is a warning rather than

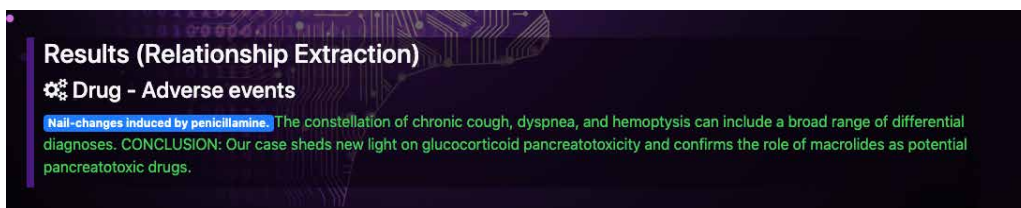
mentioned. For example, if the term 'headache' is mentioned in the same article as a drug, is it because it is a side-effect caused by that drug or the indication that the drug is treating? Unfortunately, computers can struggle to replicate this judgement process and the lack of good quality training data has limited the application of machine learning in this area.

TERMite can rapidly identify all sentences that mention adverse events in an entire corpus of scientific literature, whether that be in the correct context or not. By serving this subset of sentences to expert curators, positive and negative datasets can be produced and used directly to the deep learning algorithm BioBERT using SciBite AI.

We then compared the results of applying i) TERMite and ii) SciBite AI to the following example sentences with the goal of identifying those that describe adverse events, which are highlighted in blue.

a statement of a causal effect between the drug and the symptom. However, the trained SciBite AI model correctly differentiated between the adverse events without the need for pre-defined rules.

Example 2: '**Nail-changes** induced by penicillamine. The constellation of chronic cough, dyspnea, and hemoptysis can include a broad range of differential diagnoses. CONCLUSION: Our case sheds new light on glucocorticoid pancreatotoxicity and confirms the role of macrolides as potential pancreatotoxic drugs.'



TERMite accurately matches 'chronic cough', 'dyspnea' and 'hemoptysis' but, in this context, these terms are not being used to describe a drug-adverse event relationship. However, the trained SciBite AI model not only correctly differentiated between the adverse events but also identified the term 'nail-changes' which was not matched within SciBite's adverse event dictionary.

Through the unique and powerful combination of semantic technologies and deep learning, SciBite AI can help Pharmaceutical companies avoid the trade-off between the volume of content and the time available to review it and identify and prioritise potential adverse events from the growing corpus of scientific literature with high confidence in near-real time.

SciBite provides the tools to generate the high quality, contextualised data necessary for AI to be effective and deliver accurate and true outputs.

SciBite's award winning platform is the culmination of tens of years of experience applying Semantic Analytics to pharmaceutical data. Global pharmaceutical companies and emerging biotechs have partnered with SciBite and are leveraging our unparalleled know-how to unlock the potential of the scientific data at their disposal.

The SciBite platform is a cost-effective solution to automate the process of cleansing, normalising and annotating data. It delivers the high quality, contextualised data which is necessary for machine learning, and other applications of Artificial Intelligence, to be effective.

To learn how SciBite can unlock the potential of your data with a flexible, easy to use, accessible environment, speak to one of our experts today or email us at contact@scibite.com

THE LANGUAGE OF SCIENCE

Head Office:

SciBite Limited
BioData Innovation Centre
Wellcome Genome Campus
Hinxton, Cambridge CB10 1DR
United Kingdom

 www.scibite.com
 contact@scibite.com
 LinkedIn: SciBite
 Twitter: @SciBite
 +44 (0)1223 786 129