





## RESEARCH ARTICLE

## Semantic text mining in early drug discovery for type 2 diabetes

Lena K. Hansson<sup>1</sup> , Rasmus Borup Hansen<sup>2</sup> , Sune Pletscher-Frankild<sup>2</sup>, Rudolfs Berzins<sup>2</sup>, Daniel Hvidberg Hansen<sup>2</sup>, Dennis Madsen<sup>3</sup> , Sten B. Christensen<sup>3</sup>, Malene Revsbech Christiansen<sup>2</sup> <sup>na</sup>, Ulrika Boulund<sup>2</sup> <sup>nb</sup>, Xenia Asbæk Wolf<sup>2</sup>, Sonny Kim Kjærulff<sup>2</sup> <sup>nc</sup>, Martijn van de Bunt<sup>3</sup>, Søren Tulin<sup>3</sup>, Thomas Skøt Jensen<sup>2</sup>, Rasmus Wernersson<sup>2,4</sup> , Jan Nygaard Jensen<sup>1</sup> <sup>nd</sup>

**1** Novo Nordisk Research Centre Oxford, Novo Nordisk Ltd., Oxford, United Kingdom, **2** Intomics A/S, Kgs. Lyngby, Denmark, **3** Novo Nordisk A/S, Bagsværd, Denmark, **4** DTU Health Tech, Technical University of Denmark, Kgs. Lyngby, Denmark

 These authors contributed equally to this work.

<sup>na</sup> Current address: Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health, University of Copenhagen, Copenhagen, Denmark

<sup>nb</sup> Current address: Laboratory of Experimental Vascular Medicine, University of Amsterdam / Academic Medical Centre, Amsterdam, The Netherlands

<sup>nc</sup> Current address: Novo Nordisk A/S, Bagsværd, Denmark

<sup>nd</sup> Current address: Boehringer Ingelheim GmbH & Co. KG, Biberach, Germany

\* [rwe@intomics.com](mailto:rwe@intomics.com)


 OPEN ACCESS

**Citation:** Hansson LK, Hansen RB, Pletscher-Frankild S, Berzins R, Hansen DH, Madsen D, et al. (2020) Semantic text mining in early drug discovery for type 2 diabetes. PLoS ONE 15(6): e0233956. <https://doi.org/10.1371/journal.pone.0233956>

**Editor:** Ilya Safro, Clemson University, UNITED STATES

**Received:** August 19, 2019

**Accepted:** May 15, 2020

**Published:** June 15, 2020

**Copyright:** © 2020 Hansson et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** This work is based on processing publicly available data from PubMed. Derived data is available at: <https://doi.org/10.5281/zenodo.3603610>.

**Funding:** The work presented was sponsored by Novo Nordisk Ltd with part of it being carried out by Novo Nordisk employees and part of it by subcontracting to Intomics A/S. Novo Nordisk supplied internal data as well as taking part in the study design and overall supervision and guidance of the project, with special focus on evaluation of

## Abstract

### Background

Surveying the scientific literature is an important part of early drug discovery; and with the ever-increasing amount of biomedical publications it is imperative to focus on the most interesting articles. Here we present a project that highlights new understanding (e.g. recently discovered modes of action) and identifies potential drug targets, via a novel, data-driven text mining approach to score type 2 diabetes (T2D) relevance. We focused on monitoring trends and jumps in T2D relevance to help us be timely informed of important breakthroughs.

### Methods

We extracted over 7 million *n*-grams from PubMed abstracts and then clustered around 240,000 linked to T2D into almost 50,000 T2D relevant ‘semantic concepts’. To score papers, we weighted the concepts based on co-mentioning with core T2D proteins. A protein’s T2D relevance was determined by combining the scores of the papers mentioning it in the five preceding years. Each week all proteins were ranked according to their T2D relevance. Furthermore, the historical distribution of changes in rank from one week to the next was used to calculate the significance of a change in rank by T2D relevance for each protein.

### Results

We show that T2D relevant papers, even those not mentioning T2D explicitly, were prioritised by relevant semantic concepts. Well known T2D proteins were therefore enriched among the top scoring proteins. Our ‘high jumpers’ identified important past developments

biological relevance. Intomics was responsible for the technical implementation, statistical analysis and benchmarking of results. The decision to publish was initiated by Novo Nordisk Ltd but in agreement with Intomics A/S.

**Competing interests:** Thomas Skøt Jensen, Rasmus Wernersson, Rasmus Borup Hansen, Sune Pletscher-Frankild, Rudolfs Berzins, Daniel Hvidberg Hansen, Malene Revsbeck, Ulrika Boulund, Xenia Asbæk Wolf and Sonny Kim Kjærulff are, or were at the time, employed at Intomics A/S. Rasmus Wernersson and Thomas Skøt Jensen are both cofounders of Intomics A/S. Lena K. Hansson, Dennis Madsen, Sten B. Christiansen, Martijn van de Bunt, Søren Tulin and Jan Nygaard Jensen are, or were at the time, employed at Novo Nordisk A/S or Novo Nordisk Ltd an affiliate to Novo Nordisk A/S. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

in the apprehension of how certain key proteins relate to T2D, indicating that our method will make us aware of future breakthroughs. In summary, this project facilitated keeping up with current T2D research by repeatedly providing short lists of potential novel targets into our early drug discovery pipeline.

## Introduction

Drug discovery often involves prioritising potential drug target candidates based on a combination of researchers' knowledge, data from wet-lab experiments, and *in-silico* predictions. The researchers' knowledge typically comes from their own work as well as publicly available scientific literature. Due to recent technological advancements, publicly available knowledge bases are growing exponentially via large-scale omics, human genetics studies, imaging projects, and electronic health records [1, 2]. Thus, for the most well-studied diseases thousands of relevant articles are available. Text mining is one of the most common tools to help choosing which articles to read, and extract explicit facts, but also to combine independent pieces of knowledge, thereby prioritising resources in the early drug discovery phase.

Type 2 diabetes (T2D) is a chronic disease with about half a billion diagnosed patients, and probably the seventh leading cause of death due to its many co-morbidities (such as blindness, kidney failure, heart attacks, stroke, and lower limb amputation) [3]. T2D is a multi-factorial disease associated with many proteins (from 30 in OMIM [4], through more than 200 genetic loci [5], to over 3,000 postulated associations in OpenTargets [6]). The continuous growth of our understanding of signalling events in T2D, and therefore relevant literature, is considerable. No single person, group, or even large organisation is able to keep up with all details, but is instead more likely to catch the high impact literature only. In 2018 more than 1 million articles were indexed by PubMed, or around 20,000 articles per week, of which we estimate 1–2%, equivalent to hundreds per week, to be relevant for T2D.

A simple keyword search for T2D synonyms will of course find all known articles where the association to T2D has already been established, but may also find articles on, for example, patient care that are less interesting in a drug discovery context. Such articles will dilute the overall results and the ability of text mining to support drug target identification. On the other hand, many relevant papers regarding the complex underlying biology may not mention T2D synonyms, either because the association is not yet established, or it may be understood from the context. Such papers may investigate co-morbidities of T2D or related metabolic processes, and may therefore not be found in a keyword search.

Text mining could be described as the process of discovering, and capturing, knowledge from a large number of unstructured texts, and it is being applied to more and more problems in drug discovery [7]. Text mining tools vary from simple over-representation statistics between two sets of keywords like [8] to more complex NLP approaches like [9]. To date, most text mining tools in the biomedical field are specialised to specific tasks [10–13]. Some systems, e.g. BioReader [14], find both direct and indirect associations by considering large numbers of search terms using machine learning algorithms, and a recent application [15] applied the Word2vec [16] algorithm to materials science literature successfully. Other tools are designed to work on a general level, such as IBM Watson [17] and SciBite's TERMite [18].

In order to find associations you have to specify what you are looking for, i.e. your vocabulary, in a Named Entity Recognition (NER) step. This means that you are inherently limited to those concepts in the specified vocabulary. There has been focus on solving this out-of-

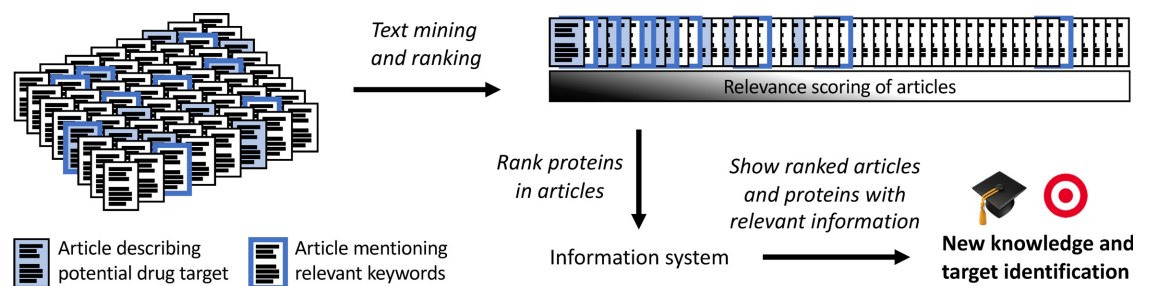
vocabulary problem for example by using machine learning to determine the vocabulary from data using  $n$ -grams [19]. Alternatively, ToPMine [20] (and tools built on this like Moliere [21]) is a purely data-driven approach which focus on identifying significant, frequent phrases.

Another approach is the ‘question/answer’ approach in which you focus on predicting the next word(s) given a certain set of preceding words [22].

There are also methods trying to find the implicit link between two concepts based on the ABC co-occurrence model, where if A and B are co-mentioned, and B and C are co-mentioned, you assume there is a link between A and C, for example [23] or similar approaches like [24]. Relying on the MeSH annotation [25] for individual papers instead of text mining, [26] studies co-annotations, aiming at forming new medical hypotheses based on evolutionary trajectories of their embeddings for different five years’ time slices.

In the present work we focus on, potentially right after publication, linking proteins and relevant mechanisms learned from the literature itself to diseases, in this case T2D, as proteins are the actual agents driving the underpinning physiological mechanisms; this is the very heart of molecular medicine. Several projects seek to associate genes/proteins to diseases; two popular, such databases, OpenTarget and DisGeNET [27] currently list thousands of T2D associated proteins. More specifically, DisGeNET currently lists 1,671 genes when querying for T2D, while Open Targets currently lists 3,346 associated genes. These numbers are growing as more knowledge is disseminated. In addition, some disease proteins yet remain to be well described in literature (see S1 File S1.1). This makes it a challenge to determine appropriate statistical criteria for disease associations. However, only a handful of proteins are currently used as targets for drugs for T2D, so we reason there is still a potential for identifying novel drug targets.

Here we present a project that addressed the problem of quantifying, and at the same time providing an explanation for, the relevance of papers for T2D. This work is being used to guide the discovery of novel T2D drug targets. Fig 1 illustrates, on a high level, how we used our text mining scores of papers’ T2D relevance to help obtain new knowledge and identify new targets. Similar to other projects, it prioritised among the thousands of proteins already tentatively linked to T2D, glucose control, and/or insulin resistance. A benefit to our approach is that we broadened the definition of the disease by learning and focusing on the underlying biological mechanisms. The broadening was done by basing our relevance score on a purely data-driven detection of multi-word terms,  $n$ -grams, and their clustering into what we call ‘semantic concepts’, rather than a set vocabulary. The molecular focus and T2D relevance was ensured by weighing the concepts according to their enrichment with articles mentioning one,



**Fig 1. Using text mining to guide early drug discovery.** Keyword searches for publications, like in PubMed, may find thousands of articles (borders highlighted), including those less relevant to drug discovery, e.g. on the care of patients. Yet, articles describing potential drug targets (background coloured) risk not being found as the disease association may be implicit or not yet established. To address these issues, articles are ranked using text mining, focusing on molecular biology without requiring direct mentioning of the disease. Proteins are then ranked, and the information system shows the articles with relevant terms highlighted in addition to pertinent information, providing a convenient workflow for advancing knowledge and identifying drug targets.

<https://doi.org/10.1371/journal.pone.0233956.g001>

or more, of the 100 proteins most significantly associated to T2D via simple synonym searches. We could then identify T2D relevant proteins, potentially even before they have been explicitly co-mentioned with T2D. Each protein was scored based on five years' worth of articles. It is well known that certain proteins have, at some points in time, started receiving more attention in the T2D relevant literature. To face the unsolved challenge of ensuring awareness of such proteins, we investigated trends and changes in T2D relevance. We introduced the term 'high-jumpers' for proteins that significantly change their rank in comparison to the distribution of rank changes seen over the previous years. Utilising existing infrastructure, the project continues to present interesting articles and proteins on a weekly basis, as it is essential for early drug discovery to be made timely aware of potential novel drug targets.

## Methods

We text mined PubMed abstracts using NER to find the proteins mentioned in each abstract. For our project we used a custom text mining engine implemented in C and Python (see [S1 File S1.1](#)), but any NER system that can find proteins mentioned in a text using a set of well-curated synonyms (e.g. PubTator [10–12], TERMite [18], or the system used at JensenLab [13]) can be used by the method we present. PubMed abstracts were also mined for mentions of T2D, and we used Fisher's exact test to measure how co-mentioned a protein was with T2D (see [S1 File S1.2](#) for details). All proteins were ranked by their *p*-values for being related to T2D (see [S1 File S1.3](#) for details), and we chose to denote the 100 most significantly associated as 'core proteins' as they have well-established and well-known associations to T2D (according to the literature captured in PubMed).

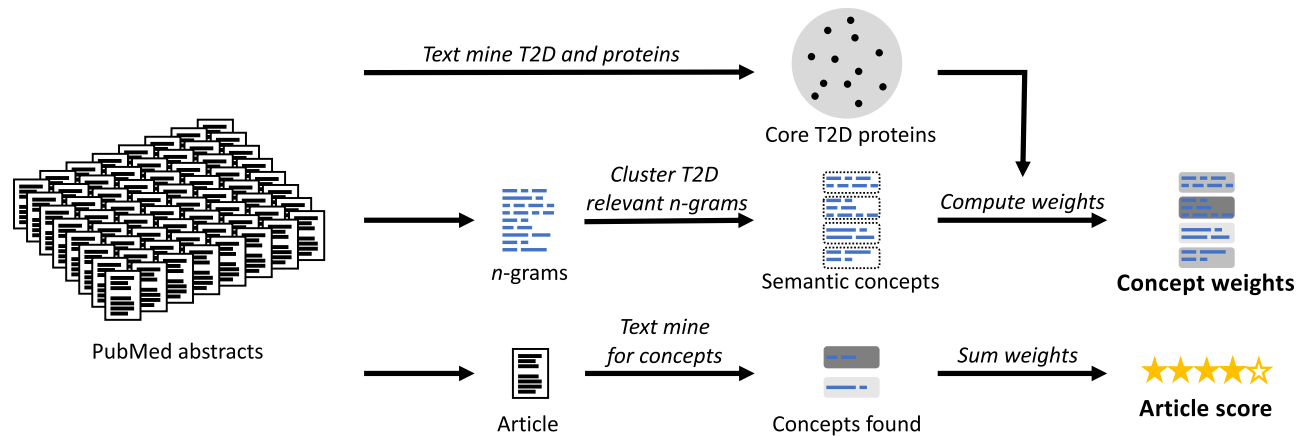
We used these 100 core proteins to characterise our biological area of interest and proceeded to construct an article scoring scheme by favouring terms co-mentioned with the 100 core proteins, and used these individual article scores to prioritise proteins according to T2D relevance.

A rigorous mathematical description of our methods and statistical analyses is found in [S1 File](#) sections S1.1 through S1.10.

## Scoring T2D relevance of articles using semantic concepts

[Fig 2](#) outlines our data-driven method for scoring article abstracts for T2D relevance. We focused on the relevant biology by revolving the analysis around the 100 core T2D proteins, and we broadened the search by looking for phrases related to the core T2D proteins rather than the disease directly. As will be further discussed under 'T2D relevance of articles' in the 'Results' section, we could in this way give high scores to articles with abstracts that did not mention T2D explicitly, but instead used many of the same phrases and terms as abstracts mentioning proteins known to be related to T2D. In addition to expanding the scope of the biological space, e.g. our understanding of how the disease works, this meant that we could potentially identify proteins before they become well known in a T2D context.

To broaden our search beyond articles mentioning T2D explicitly, and to focus on articles on molecular biology (in contrast to e.g. care of diabetes patients), we set out to find articles that discuss the same biology in their abstracts as articles mentioning the 100 core T2D proteins. For this, we needed a set of entities and synonyms that could be used to properly characterise the abstracts. Even though a large effort has been put into creating ontologies like MeSH and UMLS [28] for biomedical terms as well as repositories like Semantic MEDLINE database [29], these resources did not encompass all relevant phrases. Instead, we decided to extract a terminology directly from PubMed in a purely data-driven manner where we identified frequently used sequences of words, *n*-grams (see [S1 File S1.4](#)).



**Fig 2. Scoring T2D relevance of articles.** Named entity recognition is used to find the 100 core T2D proteins that are most significantly co-mentioned with T2D in PubMed abstracts. Commonly used word sequences ( $n$ -grams) are learned from the abstracts and T2D relevant ones are clustered into 'semantic concepts', if they are textually similar, or used in the same context. The semantic concepts are assigned weights based on how often they are co-mentioned with the core T2D proteins. All articles, old and new, are then scored by summing the weights of the semantic concepts found in them.

<https://doi.org/10.1371/journal.pone.0233956.g002>

Many  $n$ -grams, even for different sizes of  $n$ , refer to the same semantic object (e.g. 'lower glucose' and 'lowers blood sugar level'). Considering all these  $n$ -grams individually diminishes the overall statistical power. We therefore grouped similar, and related,  $n$ -grams into what we coined 'semantic concepts'. We first constructed a similarity measure to join textually similar  $n$ -grams that are spelled slightly differently, or have their words rearranged, e.g. 'type 2 diabetes' and 'diabetes type II' (see [S1 File S1.5](#) for details). Secondly, we measured the semantic relatedness between  $n$ -grams by computing how often they co-occur in relevant articles by constructing a similarity matrix where each cell contains the over-representation ratio for the two  $n$ -grams corresponding to the row and the column (see [S1 File S1.6](#)). The rationale for using this similarity matrix, was that if, say, two rows in the matrix were very similar, the two  $n$ -grams would be co-mentioned with the same set of  $n$ -grams and hence tend to be used in the same context, and would be likely to share the same 'meaning', even though they need not be synonyms (for example 'leptin treatment' and 'elevated leptin' are not synonyms but are considered part of the same semantic concept). We finally used the MCL algorithm [30] to group  $n$ -grams into semantic concepts based on these two similarity measures simultaneously (see [S1 File S1.7](#) for details).

Our semantic concepts are therefore clusters of  $n$ -grams that have both some degree of string similarity as well as similar co-mentioning with other  $n$ -grams; i.e. they behave the same way semantically and look alike string-wise. It is important to note that the individual  $n$ -grams for each semantic concept need not be synonyms, but they are expected to be equally relevant in a given context (e.g. 'increased glucose', 'decreased glucose' and 'total glucose' are clustered together by our method along with other glucose related  $n$ -grams).

We assigned weights to the semantic concepts by first computing how much more than expected by random they were co-mentioned with the core set of 100 T2D proteins. For each of the core proteins, and for each semantic concept, this ratio was log-transformed to get the what is known as the pointwise mutual information [31], and the weight for each semantic concept was computed as the average across all the core proteins (see [S1 File S1.8](#) for details). An article was scored by simply taking the sum of the individual weights of all the semantic concepts with  $n$ -grams in the article, thereby broadening our search beyond articles explicitly mentioning T2D, and at the same time zooming in on articles related to the molecular biology

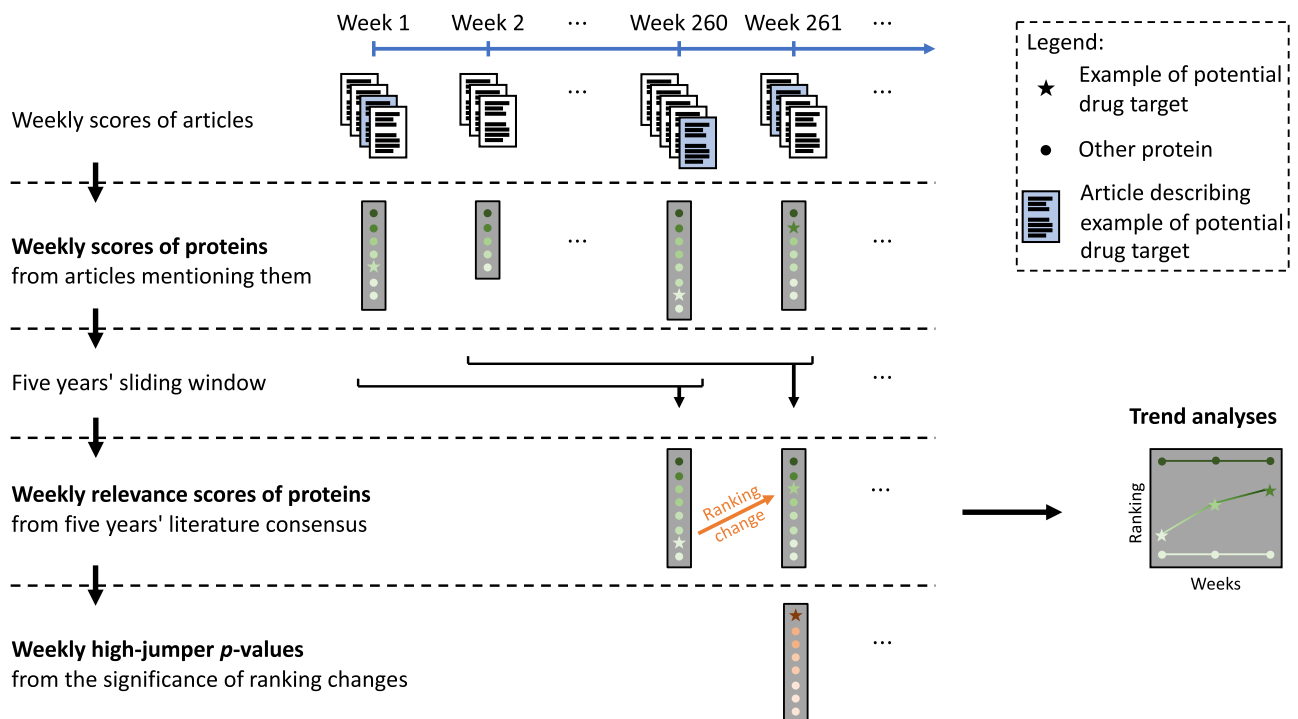
of T2D. In other words, an article will get a high score, if it uses similar words as articles mentioning the 100 core T2D proteins, and the more specific these words are, the higher the score.

### Determining T2D relevance of proteins

In this project we computed two different T2D relevance scores for proteins, both based on the scores of the articles mentioning them (see Fig 3).

**Five years' T2D relevance score.** As the weekly score of a protein we used the highest score of an article mentioning the protein in that particular week. We then ranked the previous 260 weekly protein scores in order to determine which proteins are the most relevant to T2D as per the combined recent research (in a five years' sliding window). The 40 highest were then weighted so the highest contributed most to the T2D relevance score for the protein. In this way, contributions had to be spread out over time and not all come from the same week (e.g. a single conference's proceedings). This is described more formally in [S1 File S1.9](#). By shifting the time period one week at a time, we could retrospectively visualise trends in T2D relevance for individual proteins.

**Weekly high-jumping proteins.** In order to find novel T2D proteins, we focused on the current week and identified proteins with high upward jumps and the abstracts causing these. This expanded our more traditional weekly literature survey, by selecting a different set of articles, and by focusing on the most important proteins in these (for example insulin tend to be mentioned in all articles but is not interesting from a new drug target point of view).



**Fig 3. Scoring T2D relevance of proteins.** Articles from PubMed are grouped by the week they were added to the database. The T2D relevance of proteins is prioritised in different ways: 1) the highest T2D relevance score for an article mentioning the protein and published a given week; 2) these scores for the last five years combined into one consensus score (updated weekly using a sliding window); and 3)  $p$ -values assessing a protein's change in the overall ranking, pinpointing high-jumping proteins that are suddenly mentioned in a T2D relevant context. In addition, looking at long-term trends of the combined scores may reveal interesting biology or proteins.

<https://doi.org/10.1371/journal.pone.0233956.g003>



Seeing a top ranking protein improve its rank by, say, three places is not very likely, and is certainly something we should be made aware of if it happens. It is more likely, but less interesting, for a protein further down the list to improve its rank by three places. A protein that jumps hundreds or thousands of places may be relevant, depending on how typical such jumps are.

To address this, the empirical distribution of jumps for each initial rank was used, i.e. we looked back in time and observed both the jumps and their starting ranks. When a protein's weekly rank changed, we consulted the distribution corresponding to its initial rank to get a  $p$ -value for the jump. However as not all jumps were observed, we smoothed the distributions by applying a sliding window (see [S1 File S1.10](#) for details). Hence, a jump in a protein's rank got a low  $p$ -value if the protein was unexpectedly mentioned in an article with high T2D relevance score, based on the combined previous knowledge about the protein.

### Supporting early drug discovery

On a weekly basis, all new articles abstracts from PubMed were text mined and ranked by our scoring scheme that prioritised abstracts with the same semantic concepts as those mentioning the core T2D proteins. The high ranking article abstracts were presented in the web interface where protein synonyms, diseases, drugs, and anatomical entities were highlighted in order to easier derive novel biological insight.

The same framework was also used to display the most relevant proteins for a given week, based on the protein scores described previously. In order to be kept informed of new trends in T2D research we ranked the proteins according to their weekly scores. In addition, in order to find novel targets before they are T2D associated, we also highlighted the 'high-jumping' proteins, i.e. the proteins that were suddenly ranked more T2D relevant than previously (as described in the previous section). For each featured protein we both retrieved annotations and linked to other internal information systems with pertinent details, such as if this protein is already known in a T2D context or not.

## Results

Of the 30.3 million articles registered in PubMed in November 2019, we text mined the 19.8 million (65%) with both a title and an abstract. From these abstracts we first identified and reviewed the set of the 100 most T2D associated proteins (see [S1 File S1.3](#)) that was the starting point for our main results described in the following sections.

### T2D relevance of articles

We broadened the scope of the disease biology, trying to better characterise T2D by implementing an unsupervised method of generating, and evaluating semantic concepts with thousands of these contributing towards the article scores. Among the more than 7 million  $n$ -grams we extracted (see [S1 File S1.4](#)), 235,382 were over-represented with T2D and these were clustered into 48,381 semantic concepts.

The use of unsupervised detection of semantic concepts directly from PubMed rather than using traditional vocabularies enabled us to capture terms that we might not necessarily have included otherwise. Investigation of the important concepts expanded our understanding of the disease, the mechanisms behind it, and the relevant tissues.

For each semantic concept we computed a weight that was used to score individual articles (see the section 'Scoring T2D relevance of articles using semantic concepts' in 'Methods' and [S1 File S1.8](#)). [Table 1](#) shows the top 50 semantic concepts we found (and their assigned weights), each represented by their most common  $n$ -gram. They all had weights close to 1,

**Table 1. Top 50 semantic concepts for T2D.** The semantic concepts with the highest weights for T2D are focused on glucose homeostasis or tolerance as well as insulin resistance, sensitivity, levels, secretion, and the insulin receptor substrate 1 before the db/db mouse model and T2D itself.

Rank	Most common <i>n</i> -gram	Weight	Abstracts	Rank	Most common <i>n</i> -gram	Weight	Abstracts
1	glucose homeostasis	1.208	8,916	26	fasting insulin	0.975	6,865
2	insulin resistance	1.155	59,462	27	leptin levels	0.962	20,700
3	insulin sensitivity	1.132	21,599	28	insulin glucose	0.961	2,550
4	aims hypothesis	1.130	5,728	29	plasma glucose concentration	0.961	5,093
5	glucose tolerance	1.121	16,714	30	diet induced obese	0.960	1,287
6	insulin receptor substrate 1	1.113	2,939	31	obese diabetic mice	0.960	611
7	insulin secretion	1.104	17,655	32	phase insulin secretion	0.959	1,743
8	improved glucose tolerance	1.099	2,330	33	improve insulin resistance	0.958	1,034
9	insulin levels	1.093	10,194	34	glucose tolerance tests	0.957	2,366
10	db db mice	1.092	4,104	35	cell counting kit	0.957	16,141
11	type 2 diabetes mellitus t2dm	1.088	13,215	36	oral glucose challenge	0.955	1,583
12	increased insulin sensitivity	1.074	2,475	37	diet induced	0.946	7,837
13	insulin resistant	1.071	5,928	38	obese mice	0.943	5,735
14	type 2 diabetes	1.070	91,290	39	diet induced changes	0.940	1,895
15	glucose metabolism	1.046	26,156	40	increased insulin	0.938	1,098
16	impaired glucose tolerance	1.041	20,710	41	plasma insulin levels	0.937	5,026
17	high fat diet	1.038	16,674	42	glucose stimulated insulin secretion	0.937	3,029
18	$\beta$ cell function	1.027	2,591	43	insulin resistance ir	0.935	5,129
19	diabetes t2d	1.022	12,100	44	insulin receptor substrate	0.934	1,589
20	plasma glucose	1.017	13,480	45	insulin glucagon	0.931	5,633
21	induced insulin resistance	1.005	2,432	46	insulin tolerance test	0.930	1,393
22	beta cells	0.997	20,780	47	insulin receptor	0.929	9,195
23	high fat diet hfd	0.993	9,226	48	homeostasis model assessment	0.929	6,774
24	ob ob mice	0.987	4,394	49	lipid metabolism	0.928	31,435
25	plasma insulin	0.984	8,133	50	insulin resistance homa ir	0.927	5,059

<https://doi.org/10.1371/journal.pone.0233956.t001>

meaning that they were on average co-mentioned 10 times as often as expected by random with one of the top 100 core proteins. The highest scoring semantic concept is a cluster of *n*-grams about glucose metabolism. The most frequently used *n*-gram in this concept is simply: ‘glucose homeostatis’ but since the literature has a natural focus on describing and understanding disease-states, the other *n*-grams reference dysfunction of this metabolic process, hence the concept also includes ‘disturb glucose homeostasis’, ‘disturbed glucose homeostasis’, ‘defective glucose metabolism’, ‘impair glucose homeostasis’, and ‘impairs glucose homeostasis’.

We manually inspected the 1,000 highest scoring article abstracts (since 2014) to assess their relevance for T2D. Of these, 406 mentioned T2D explicitly, 181 had the MeSH term ‘Diabetes Mellitus, Type 2’ (with 168 overlapping). The remaining 581 abstracts were often about insulin resistance, or a well-known co-morbidity such as obesity, non-alcoholic fatty liver disease, diabetic neuropathy, kidney, or cardiovascular disease. Another prevalent theme among the inspected abstracts is that they talk about genes (variation, expression or knockout) or experiments in animals. Occasionally, gestational diabetes, type 1 diabetes, or streptozotocin induced diabetes would instead be the focus on the abstract.

We also inspected abstracts with the ‘Diabetes Mellitus, Type 2’ MeSH term, but with low scores, and found that these articles were often about case reports, the care of T2D patients, or it was not evident from the title or abstract alone why the article was tagged with the MeSH term for T2D, for example [32] which discusses an extension of a patient questionnaire that was then evaluated on T2D patients.

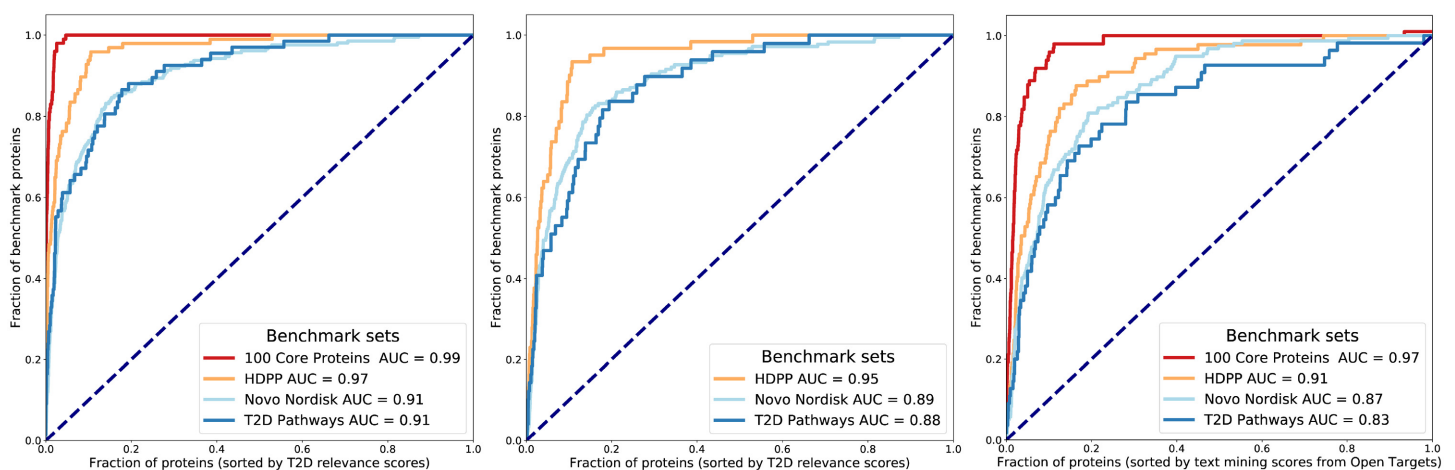


As can be seen in Table 1 the system separated semantic concepts of a similar nature like ‘type 2 diabetes’, ‘type 2 diabetes mellitus t2dm’, and ‘diabetes t2d’; even though these would normally be considered synonyms for the same biological concept. While this may at first seem surprising, we do note that these concepts get similar, high weights. In general, the data-driven approach generated higher resolution concepts compared to manually maintained ontologies, enabling us to separate terms and thereby gave us a greater understanding of the biological space. For example, ‘glucose tolerance’ had the 5th highest weight, ‘improved glucose tolerance’ had the 8th highest, while ‘normal glucose tolerance’ had the 54th highest weight (not shown).

As a curiosity we noted that the semantic concept with the 4th highest weight contained the  $n$ -gram ‘aims hypothesis’ because abstracts from the journal *Diabetologia* all have the heading ‘aims/hypothesis’. Also, the 35th highest ranking semantic concept is represented by the  $n$ -gram ‘cell counting kit’ in the table as this is the most prevalent  $n$ -gram for the semantic concept that also encompasses more than 100 different  $n$ -grams including ‘pancreatic  $\beta$  cells’. Apart from these two cases, all the semantic concepts with the highest weights are, as expected, highly relevant for T2D (or obesity).

## T2D relevance of proteins

As we computed T2D relevance scores for proteins from the last five years’ of literature, we expected well-known T2D proteins to get the highest scores. We benchmarked this by examining how well the scoring scheme for this project was able to find three sets of T2D proteins via ROC curves; 1) T2D proteins from the Human Diabetes Proteome Project [33], 2) a manual benchmark list of proteins selected for this project, and 3) proteins from public T2D pathways (KEGG [34–36], WikiPathways [37]) or annotated in UniProt [38] as relevant for T2D (Fig 4, left and centre). All had AUCs above 0.88, i.e. using our scoring scheme, the set of high scoring proteins was clearly enriched for the well known T2D proteins. We also compared our scoring to the text mining scores from Open Targets [39] (Fig 4, right) and noted that we have slightly higher AUCs.



**Fig 4. Benchmarking the T2D protein relevance scores.** *Left:* The ROC curves show that we find a) the 100 core proteins used to characterise our area of interest (with an AUC very close to 1), b) the 100 diabetes relevant proteins identified by the HDPP project (3 of these proteins are not found by text mining), c) a set of 209 proteins of particular interest to the project, and d) 66 proteins from public type 2 diabetes pathways (KEGG and WikiPathways) or annotated in UniProtKB as relevant for T2D. *Centre:* Here, the 100 core T2D proteins are removed from each benchmark set (62 remain in HDPP, 181 in the internal set, and 47 in the T2D pathways), which obviously lead to slight AUC reductions. *Right:* The curves when using text mining scores from Open Targets [39].

<https://doi.org/10.1371/journal.pone.0233956.g004>

**Table 2. Top 100 most T2D relevant proteins from the last 5 years according to our scoring scheme, sorted alphabetically by gene name.** The 54 proteins with an asterisk (\*) are also part of the set of 100 core T2D proteins.

ACOX1	BCL2	DECR1	GCGR*	HSD11B1*	ITGAX	NAMPT*	PNPLA2	RBP4*	SOCS3
ADIPOQ*	CASP3	DGAT1	GCK*	HSPA5	LDLR	NFE2L2	POMC	REN*	SREBF1*
ADIPOR1*	CAT	DPP4*	GHRL*	ICAM1	LEP*	NFKBIA	PPARA*	RETN*	SREBF2
AGT*	CCK	EEF1A2*	GIPR*	IGF1*	LEPR*	NOS2	PPARD*	SCD	STAT3
AKT1*	CCL2*	FABP4*	GLP1R*	IL10	LPL*	NOS3*	PPARG*	SERPINE1*	STK11
AKT2	CD36*	FASN	GOT2	IL1B	MAPK3	NOX4	PPARGC1A*	SIRT1	TGFB1
APOA1*	CD68	FGF21*	GPX1	IL6*	MAPK8	NR1H3	PRKCE	SLC2A1*	TLR4
APOB*	CEBPA	FOXO1*	GSK3B*	INS*	MGAM*	PCK1	PTEN	SLC2A2*	TNF*
APOC3*	CPT2	G6PC*	HMGCR	INSR*	MLXIPL	PDK4	PTGS2	SLC2A4*	UCP1*
APOE*	CRP*	GCG*	HMOX1	IRS1*	MTOR	PDX1*	PTPN1*	SLC5A2*	VCAM1*

<https://doi.org/10.1371/journal.pone.0233956.t002>

Table 2 shows the 100 proteins with the highest relevance scores. There is, not surprisingly, a large overlap with the 100 core proteins. The proteins, and papers, that we had not previously flagged were investigated, and almost all of them were deemed interesting. This shows that this automatic data-driven approach works, as it keeps the focus on T2D while at the same time expanding the bio-molecular knowledge space around T2D.

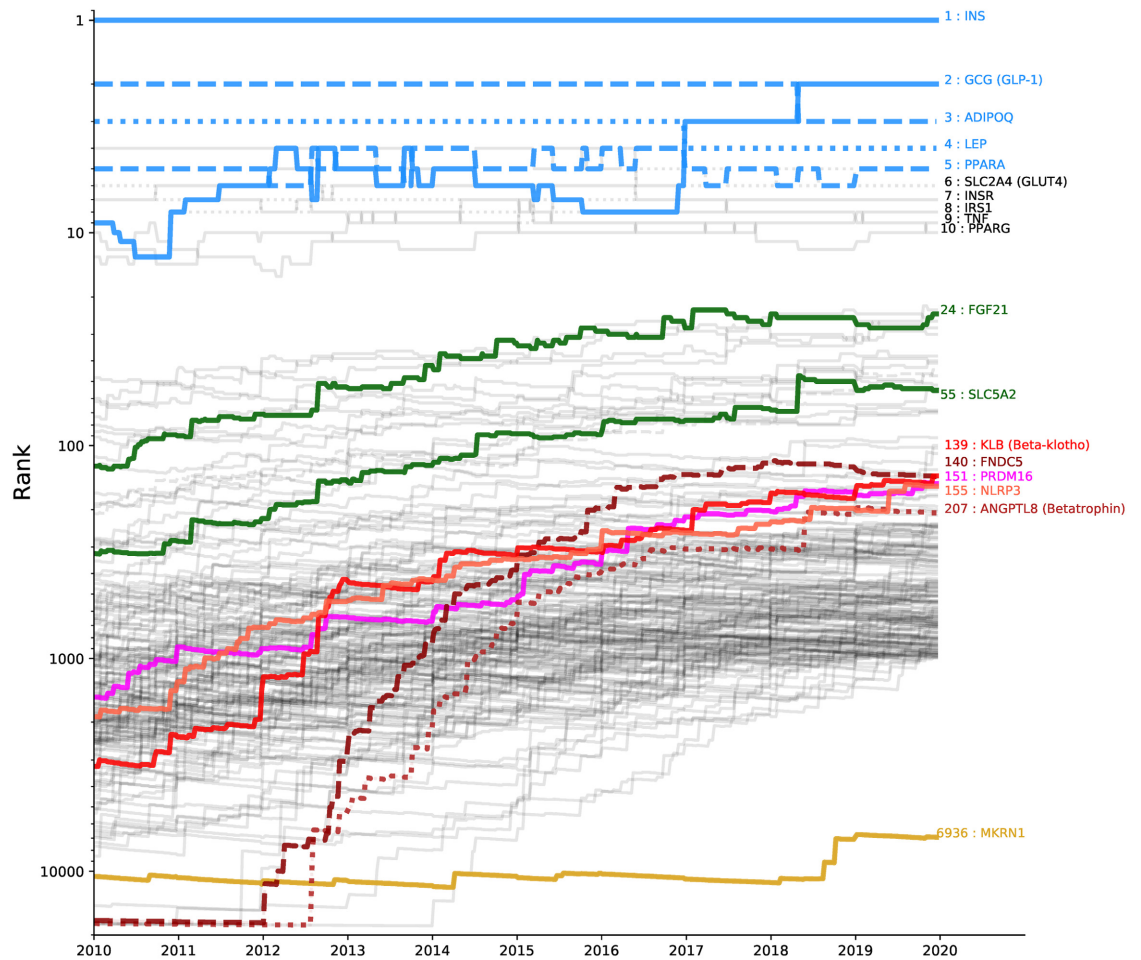
**Trends in T2D relevance.** It is known that pharmacology for T2D has advanced rapidly during the last 10 years and it is now possible to target different pathophysiological defects (as the existing classes of T2D drugs do). Accordingly, we observed that for certain proteins the strength of their association to T2D change over time; e.g. going from the initial discovery to a more detailed understanding of the protein's biological role. To study this evolution, we plotted the protein ranks as they changed over the last 10 years. Fig 5 shows these curves for the proteins that improved their ranks two-fold, or more, over the period plus the 10 most associated proteins. We found these trend curves to be generally representative of the development in T2D research, as we will show in the following sections by highlighting a few examples.

As expected, insulin is constantly the top ranking T2D protein (it's effective in all disease stages and necessary to achieve glycemic control [40]). The second place belonged to adiponectin for more than 8 years. This protein is secreted by adipose tissue and is a metabolic regulator of glucose control and breakdown of fatty acids [41]. In addition, adiponectin can reverse insulin resistance in mice in combination with leptin [42]. Adiponectin is therefore strongly involved in T2D processes. In the beginning of 2018 its second place was taken by glucagon (GLP-1).

The high ranking of leptin can be explained by the fact that it was discovered as the hormone regulating appetite in obese mice, which is a known co-morbidity to T2D, but also raised considerable interests within the T2D research community for its possible implication in sustaining insulin resistance [43] and ability to positively influence plasma glucose levels in diabetic mice [44]. Leptin was a clear third from the beginning of our trend analysis until the end of 2016 where glucagon (GLP-1) caught up.

The 5th place is PPARA (Peroxisome proliferator-activated receptor alpha) which since 2000 has been known to improve insulin sensitivity [45]. Eleven genetic variances (11 SNPs) were associated with response to an anti-diabetic drug (acarbose) in 2007 [46]. The recent development around PPARA is in large part due to [47], as well as [48] and [49].

On a global level we saw considerable more movement than we expected within the 100 most associated T2D proteins. To illustrate this, we overlaid the 2 proteins that improved their rankings at least 5 times and ended up among the top 100 (green curves, Fig 5). For example, FGF21 (fibroblast growth factor 21) is implicated in uptake of glucose in adipose tissue, but



**Fig 5. Evolution and jumps of protein rankings.** For a selection of proteins the curves show the ranks during the last 10 years. We include the top 10 proteins at the time of our analysis, where the top 5 are coloured in blue. Additionally, 180 proteins that improved their ranks with a factor of at least 2 are included. The green curves show FGF21 and SLC5A2, the 2 proteins among the top 100 which improved their ranks 5 times, or more. The redish curves show proteins improving ranks at least 10 times, ending up among the top 300 proteins: KLB (beta-klotho), FNDC5 (dashed), PRDM16, NLRP3 and ANGPTL8 (betatrophin; dotted). In addition, the golden curve near the bottom shows MKRN1, an example that has 2 significant weekly jumps.

<https://doi.org/10.1371/journal.pone.0233956.g005>

not in other tissues [50], it had an initial rank of 125 at the beginning of 2010, yet managed to improve its rank 5.2 times ending up at rank 24 by the end of 2019. Another significant mover that came from well outside top 100 (place 322) but settled within top 100, at place 55, was SLC5A2, or sodium/glucose cotransporter 2, responsible for retention of blood glucose by re-absorption of glucose in the kidneys [51].

**Detecting jumps in T2D relevance.** Although many likely T2D targets are part of the top scoring proteins, they do not represent a source of novel disease associations. Thus, we were also interested in suggesting articles that either establish a ‘first’ strong connection to the disease or greatly improve an already existing weak association. We expected traditional text mining approaches based solely on numbers of co-mentionings to only find well-studied proteins with obvious disease associations, as novel proteins would not have sufficiently many mentionings to be found. Thus, to allow novel proteins and valuable research findings to be uncovered, even with few mentionings, we investigated if a protein was suddenly mentioned in a T2D relevant context.

To investigate how proteins changed their T2D rank each week, and to study when the most significant jumps occur, we computed the empirical distribution of jumps given the initial rank of a protein and then investigated the most extreme changes. In general, most proteins retained their rank from week to week, and hence the median change is very close to zero. This is the result of either no new published associations, or because there were no papers with large enough scores to improve the overall protein rank. Only a few proteins compete for a top placement and most are irrelevant for early drug discovery since they have already been examined as potential drug-targets. When proteins do change rank, the typical jump was  $-1$  caused by a large positive jump of another protein. An interesting minority of changes were jumps for proteins jumping hundreds of places, signifying a specific publication with a large impact towards our understanding of the protein's function in T2D context.

Incidentally, angiopoietin-like 8 (ANGPTL8), previously known as RIFL or betatrophin, was identified as an 'insulin target' in 2012 by [52] causing a jump of 8,645 places from rank 17,955 to 9,310, corresponding to a  $p$ -value of 0.000075 (the first jump for the dotted, dark red curve in Fig 5). It was considered a potential drug-target due to its alleged role as a pancreatic  $\beta$  cell peptide growth hormone [53]. Although this claim was quickly retracted [54] as other groups failed to replicate the finding, research still continued into ANGPTL8's function due to its negative correlation with improved fasting glucose levels [55], [56] and its ability to regulate serum triglyceride levels [57]). This combined research led to a total placement in the top 400 in 2018 when it was discovered that disinhibition of adipose tissue lipoprotein lipase is a novel therapeutic modality of ANGPTL8, to enhance adipose lipid uptake and treat non-alcoholic fatty liver disease and insulin resistance [58]. This finding (see Fig 6) caused another significant jump in rank corresponding to the almost vertical, dashed, red line in mid 2018 in Fig 5.

### Angptl8 antisense oligonucleotide improves adipose lipid metabolism and prevents diet-induced NAFLD and hepatic insulin resistance in rodents.

Vatner DF<sup>1</sup>, Goedeke L<sup>2</sup>, Camporez JG<sup>2</sup>, Lyu K<sup>3</sup>, Nasiri AR<sup>2</sup>, Zhang D<sup>4</sup>, Bhanot S<sup>5</sup>, Murray SF<sup>5</sup>, Still CD<sup>6</sup>, Gerhard GS<sup>6</sup>, Shulman GI<sup>2,3,4</sup>, Samuel VT<sup>7,8</sup>.

High (W=1.17)

#### Author information

#### Abstract

**AIMS/HYPOTHESIS:** Targeting regulators of adipose tissue lipoprotein lipase could enhance adipose lipid clearance, prevent ectopic lipid accumulation and consequently ameliorate insulin resistance and type 2 diabetes. Angiopoietin-like 8 (ANGPTL8) is an insulin-regulated lipoprotein lipase inhibitor strongly expressed in murine adipose tissue. However, Angptl8 knockout mice do not have improved insulin resistance. We hypothesised that pharmacological inhibition, using a second-generation antisense oligonucleotide (ASO) against Angptl8 in adult high-fat-fed rodents, would prevent ectopic lipid accumulation and insulin resistance by promoting adipose lipid uptake.

**METHODS:** ANGPTL8 expression was assessed by quantitative PCR in omental adipose tissue of bariatric surgery patients. High-fat-fed Sprague Dawley rats and C57BL/6 mice were treated with ASO against Angptl8 and insulin sensitivity was assessed by hyperinsulinaemic-euglycaemic clamps in rats and glucose tolerance tests in mice. Factors mediating lipid-induced hepatic insulin resistance were assessed, including lipid content, protein kinase C $\epsilon$  (PKC $\epsilon$ ) activation and insulin-stimulated Akt phosphorylation. Rat adipose lipid uptake was assessed by mixed meal tolerance tests. Murine energy balance was assessed by indirect calorimetry.

**RESULTS:** Omental fat ANGPTL8 mRNA expression is higher in obese individuals with fatty liver and insulin resistance compared with BMI-matched insulin-sensitive individuals. Angptl8 ASO prevented hepatic steatosis, PKC $\epsilon$  activation and hepatic insulin resistance in high-fat-fed rats. Postprandial triacylglycerol uptake in white adipose tissue was increased in Angptl8 ASO-treated rats. Angptl8 ASO protected high-fat-fed mice from glucose intolerance. Although there was no change in net energy balance, Angptl8 ASO increased fat mass in high-fat-fed mice.

**CONCLUSIONS/INTERPRETATION:** Disinhibition of adipose tissue lipoprotein lipase is a novel therapeutic modality to enhance adipose lipid uptake and treat non-alcoholic fatty liver disease and insulin resistance. In line with this, adipose ANGPTL8 is a candidate therapeutic target for these conditions.

Low (W=0.04)  
W – item weight

**Fig 6. An example of a high scoring article causing a significant jump.** This paper on angiopoietin-like 8 caused the rank of ANGPTL8 to jump up 107 places from rank 312 to 205, corresponding to a  $p$ -value of 0.000038 in week 22 of 2018. The text is reprinted from the abstract of [58], licensed under the Creative Commons Attribution 4.0 International License [59].

<https://doi.org/10.1371/journal.pone.0233956.g006>



Nat Commun. 2018 Aug 24;9(1):3404. doi: 10.1038/s41467-018-05721-4.

### Loss of the E3 ubiquitin ligase MKRN1 represses diet-induced metabolic syndrome through AMPK activation. High (W=0.93)

Lee MS<sup>1</sup>, Han HJ<sup>2</sup>, Han SY<sup>2</sup>, Kim IY<sup>3,4</sup>, Chae S<sup>5</sup>, Lee CS<sup>2</sup>, Kim SE<sup>2</sup>, Yoon SG<sup>4</sup>, Park JW<sup>4</sup>, Kim JH<sup>2</sup>, Shin S<sup>2</sup>, Jeong M<sup>2</sup>, Ko A<sup>2</sup>, Lee HY<sup>6</sup>, Oh KJ<sup>7</sup>, Lee YH<sup>8</sup>, Bae KH<sup>7</sup>, Koo SH<sup>9</sup>, Kim JW<sup>10</sup>, Seong JK<sup>3,4</sup>, Hwang D<sup>5</sup>, Song J<sup>11</sup>.

#### Author information

#### Abstract

AMP-activated protein kinase (AMPK) plays a key role in controlling energy metabolism in response to physiological and nutritional status. Although AMPK activation has been proposed as a promising molecular target for treating obesity and its related comorbidities, the use of pharmacological AMPK activators has been met with contradictory therapeutic challenges. Here we show a regulatory mechanism for AMPK through its ubiquitination and degradation by the E3 ubiquitin ligase makorin ring finger protein 1 (MKRN1). MKRN1 depletion promotes glucose consumption and suppresses lipid accumulation due to AMPK stabilisation and activation. Accordingly, MKRN1-null mice show chronic AMPK activation in both liver and adipose tissue, resulting in significant suppression of diet-induced metabolic syndrome. We demonstrate also its therapeutic effect by administering shRNA targeting MKRN1 into obese mice that reverses non-alcoholic fatty liver disease. We suggest that ubiquitin-dependent AMPK degradation represents a target therapeutic strategy for metabolic disorders.

Low (W=0.20)  
W - item weight

**Fig 7. An example of a significant high-jumping article.** This single article caused MKRN1 to jump from rank 10,853 to 9,082 with a  $p$ -value of 0.00087 in week 34 of 2018. The text is reprinted from the abstract of [60], licensed under the Creative Commons Attribution 4.0 International License [59].

<https://doi.org/10.1371/journal.pone.0233956.g007>

Another example of an interesting high-jumping event is an article that does not mention T2D explicitly, but argues that AMPK degradation represents a therapeutic strategy for metabolic disorders. It describes an observation where the deletion of MKRN1 promotes glucose consumption and suppress lipid accumulation. The main focus of the abstract is on obesity and its co-morbidities, such as nonalcoholic steatohepatitis (NASH) [60]. The article's score is low enough that it was only the 55th most T2D relevant paper in week 34 of 2018, so without the high-jumper analysis we could have missed it (Fig 7 depicts the abstract text, its semantic concepts, and their corresponding weights).

This article corresponds to the first ( $p = 0.00087$ ) of the two significant, golden coloured jumps seen in the second half of 2018 Fig 5. The second jump ( $p = 0.00019$ ) was for the 11th most T2D relevant article in week 41 which concludes with 'implicating MKRN1 as a possible therapeutic target for metabolic syndromes, such as obesity, type II diabetes, and fat liver diseases' [61]. Therefore our high-jumping analysis correctly suggested MKRN1 as a new T2D target before it was explicitly stated, and before the suggestion that glycemic regulation by AMPK could be a therapy for T2D [62].

## Discussion

By integrating this project into the pre-existing, weekly literature surveys we made it easier to investigate the literature in multiple ways, thereby accelerating and empowering the early drug discovery process for T2D.

Our data-driven, automatic scoring of papers using text mining guided researchers to papers they may not otherwise have been aware of. Automatic detecting and weighting  $n$ -grams prioritised papers that did not mention T2D explicitly, but rather focused on the molecular biology, as we built our scoring scheme around the 100 most significantly disease associated proteins. In particular, we found and used  $n$ -grams that were not present in relevant dictionaries. A possible limitation of this part of our method is that the clustering of  $n$ -grams needs to be done for each disease individually, as only  $n$ -grams co-mentioned with T2D were considered. We believe however, that we can further enhance our method and get a set of universal semantic concepts. In addition, the grouping of  $n$ -grams into semantic concepts

occasionally led to some artefacts, such as splitting synonyms across multiple concepts, or grouping  $n$ -grams that were not equivalent. Nonetheless, we believe we obtained a more fine-grained understanding of the underlying biological processes from investigating the semantic concepts separating for instance the positive, negative, and neutral regulation of a specific function. A benefit of our scoring scheme was that it is easily explainable as it sums the concepts' weights, each of which can be visualised as per e.g. [Fig 7](#).

The article scoring scheme we introduced here aims at giving high scores to papers leading up to establishing new associations between a protein and T2D biology. These individual article scores were turned into weekly protein scores, that were investigated across a five years' window, and we have shown that this integrated score could identify the well-known T2D proteins with a slightly better performance than with the text mining scores from Open Targets. By focusing on different sliding windows we found well-known past changes of direction in T2D research. While we have thoroughly vetted the NER for the 100 T2D associated proteins our scoring scheme is based on, the possibility of false positive NER results represents a limitation to our method when identifying proteins. We faced this by incorporating NER results and markup in our framework for displaying the outcome of our analysis, making false positives easier to discern.

Our relevance scoring of proteins, and subsequent identification of high-jumping proteins, overcomes the inherent issue in protein-centric text mining that the best characterised proteins typically dominate the results. It is a major strength of our method that we can detect T2D relevance in single papers, even for proteins with no strong, prior association to T2D. These novel proteins are more intriguing from a target identification point of view, but at the same time require more validations. Typically, we investigated a handful of high-jumping proteins each week. As we re-discovered previous case-stories we believe we have a useful surveillance mechanism in place likely to make us aware of scientific breakthroughs early on.

The fact that we were encouraged to extend the project to cover other disease areas is perhaps the best indication that it contributes positively to the early drug discovery process.

## Supporting information

**S1 File.**  
(PDF)

## Acknowledgments

The authors would like to thank Luis Ayala for help with preparing the figures.

## Author Contributions

**Conceptualization:** Daniel Hvidberg Hansen, Thomas Skøt Jensen, Rasmus Wernersson, Jan Nygaard Jensen.

**Data curation:** Lena K. Hansson, Rasmus Borup Hansen, Sune Pletscher-Frankild, Rudolfs Berzins, Malene Revsbech Christiansen, Ulrika Boulund, Xenia Asbæk Wolf.

**Funding acquisition:** Søren Tulin, Jan Nygaard Jensen.

**Investigation:** Lena K. Hansson, Rasmus Borup Hansen, Sune Pletscher-Frankild, Rudolfs Berzins, Daniel Hvidberg Hansen, Malene Revsbech Christiansen, Xenia Asbæk Wolf, Rasmus Wernersson, Jan Nygaard Jensen.

**Methodology:** Lena K. Hansson, Rasmus Borup Hansen, Sune Pletscher-Frankild, Daniel Hvidberg Hansen, Dennis Madsen, Martijn van de Bunt, Rasmus Wernersson.



**Project administration:** Lena K. Hansson, Rasmus Borup Hansen, Jan Nygaard Jensen.

**Resources:** Jan Nygaard Jensen.

**Software:** Lena K. Hansson, Rasmus Borup Hansen, Sune Pletscher-Frankild, Rudolfs Berzins, Daniel Hvidberg Hansen, Sonny Kim Kjærulff, Rasmus Wernersson.

**Supervision:** Rasmus Borup Hansen, Dennis Madsen, Sten B. Christensen, Martijn van de Bunt, Thomas Skøt Jensen, Rasmus Wernersson, Jan Nygaard Jensen.

**Visualization:** Rudolfs Berzins, Dennis Madsen, Sonny Kim Kjærulff.

**Writing – original draft:** Lena K. Hansson, Rasmus Borup Hansen, Sune Pletscher-Frankild, Rudolfs Berzins.

**Writing – review & editing:** Lena K. Hansson, Rasmus Borup Hansen, Sune Pletscher-Frankild, Daniel Hvidberg Hansen, Xenia Asbæk Wolf, Rasmus Wernersson, Jan Nygaard Jensen.

## References

1. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet.* 2012; 13(6):395–405. <https://doi.org/10.1038/nrg3208>
2. Eriksson R, Werge T, Jensen LJ, Brunak S. Dose-specific adverse drug reaction identification in electronic patient records: temporal data mining in an inpatient psychiatric population. *Drug Saf.* 2014; 37(4):237–247. <https://doi.org/10.1007/s40264-014-0145-z>
3. World Health Organisation fact sheet on diabetes. Available from: <https://who.int/news-room/fact-sheets/detail/diabetes>. [Accessed 14 August 2019].
4. Online Mendelian Inheritance in Man (OMIM) search for type 2 diabetes. Available from: <https://www.omim.org/entry/125853?search=type%20%20diabetes>. [Accessed 29 December 2019].
5. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature Genetics.* 2018; 50:1505–1513. <https://doi.org/10.1038/s41588-018-0241-6> PMID: 30297969
6. Open Targets Platform associations for type II diabetes mellitus. Available from: [https://www.targetvalidation.org/disease/EFO\\_0001360/associations](https://www.targetvalidation.org/disease/EFO_0001360/associations). [Accessed 29 December 2019].
7. Zheng S, Dharssi S, Wu M, Li J, Lu Z. Text Mining for Drug Discovery. *Methods Mol Biol.* 2019; 1939:231–252. [https://doi.org/10.1007/978-1-4939-9089-4\\_13](https://doi.org/10.1007/978-1-4939-9089-4_13)
8. Kuusisto F, Steill J, Kuang Z, Thomson J, Page D, Stewart R. A Simple Text Mining Approach for Ranking Pairwise Associations in Biomedical Applications. *AMIA Jt Summits Transl Sci Proc.* 2017; 2017:166–174.
9. Kogan E, Twyman K, Heap J, Milentijevic D, Lin JH, Alberts M. Assessing stroke severity using electronic health record data: a machine learning approach. *BMC Med Inform Decis Mak.* 2020; 20(1):8. <https://doi.org/10.1186/s12911-019-1010-x>
10. Wei CH, Kao HY, Lu Z. PubTator: a Web-based text mining tool for assisting Biocuration. *Nucleic Acids Research.* 2013; 41. <https://doi.org/10.1093/nar/gkt441>
11. Wei CH, Harris BR, Li D, Berardini TZ, Huala E, Kao HY, et al. Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts. *Database(oxford).* 2012; 18.
12. Wei CH, Kao HY, Lu Z. PubTator: A PubMed-like interactive curation system for document triage and literature curation. *BioCreative 2012 workshop.* 2012; 05.
13. Pletscher-Frankild S, Palleja A, Tsafou K, Binder JX, Jensen LJ. DISEASES: text mining and data integration of disease-gene associations. *Methods.* 2015; 74:83–89. <https://doi.org/10.1016/j.ymeth.2014.11.020>
14. Simon C, Davidsen K, Hansen C, Seymour E, Barnkob MB, Olsen LR. BioReader: a text mining tool for performing classification of biomedical literature. *BMC Bioinformatics.* 2019; 19(13):57. <https://doi.org/10.1186/s12859-019-2607-x>
15. Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature.* 2019; 571(7763):95–98. <https://doi.org/10.1038/s41586-019-1335-8> PMID: 31270483

16. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. In: Bengio Y, LeCun Y, editors. 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings; 2013. Available from: <http://arxiv.org/abs/1301.3781>.
17. Homepage for IBM Watson. Available from: <https://www.ibm.com/watson>. [Accessed 9 January 2020].
18. Homepage for Scibite: TERMite. Available from: <https://www.scibite.com/platform/termite/>. [Accessed 21 February 2019].
19. Cho H, Lee H. Biomedical named entity recognition using deep neural networks with contextual information. *BMC Bioinformatics*. 2019; 20(1):735. <https://doi.org/10.1186/s12859-019-3321-4>
20. El-Kishky A, Song Y, Wang C, Voss CR, Han J. Scalable Topical Phrase Mining from Text Corpora. *Proc VLDB Endow*. 2014; 8(3):305–316. <https://doi.org/10.14778/2735508.2735519>
21. Sybrandt J, Shtutman M, Safro I. MOLIERE: Automatic Biomedical Hypothesis Generation System. *KDD*. 2017; 2017:1633–1642.
22. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: NAACL-HLT; 2019. Available from: <https://arxiv.org/pdf/1810.04805.pdf>.
23. Henry S, Panahi A, Wijesinghe DS, McInnes BT. A Literature Based Discovery Visualization System with Hierarchical Clustering and Linking Set Associations. *AMIA Jt Summits Transl Sci Proc*. 2019; 2019:582–591.
24. Swanson D, Neii, Smaheiser R. Implicit text linkages between Medline records: using Arrowsmith as an aid to scientific discovery. *Library Trends*. 1999; p. 48–59.
25. Homepage for Medical Subject Headings. Available from: <https://www.nlm.nih.gov/mesh/>. [Accessed 21 February 2019].
26. Xun G, Jha K, Gopalakrishnan V, Li Y, Zhang A. Generating Medical Hypotheses Based on Evolutionary Medical Concepts. In: 2017 IEEE International Conference on Data Mining (ICDM); 2017. p. 535–544. Available from: <https://doi.org/10.1109/ICDM.2017.63>.
27. DisGeNET search for Diabetes Mellitus, Non-Insulin-Dependent. Available from: <http://www.disgenet.org/browser/0/1/0/C0011860/>. [Accessed 21 February 2019].
28. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. 2004; 32(Database issue):D267–270. <https://doi.org/10.1093/nar/gkh061>
29. Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindfleisch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics*. 2012; 28(23):3158–3160. <https://doi.org/10.1093/bioinformatics/bts591>
30. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*. 2002; 30(7):1575–1584. <https://doi.org/10.1093/nar/30.7.1575>
31. Fano RM. Transmission of Information: A Statistical Theory of Communications. *American Journal of Physics*. 1961; 29(11):793–794. <https://doi.org/10.1119/1.1937609>
32. Alvarado-Villa GE, Moncayo-Rizzo JD, Gallardo-Rumbela JA. Spanish validation endorsement of SDM-Q-9, a new approach. *BMC Public Health*. 2019; 19(1):106. <https://doi.org/10.1186/s12889-019-6436-7>
33. Schvartz D, Bergsten P, Baek KH, Barba De La Rosa A, Cantley J, Dayon L, et al. The Human Diabetes Proteome Project (HDPP): The 2014 update. *Translational Proteomics*. 2015; 8-9:1–7. <https://doi.org/10.1016/j.tprot.2015.03.001>
34. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res*. 2019; 47(D1):D590–D595. <https://doi.org/10.1093/nar/gky962>
35. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. 2017; 45(D1):D353–D361. <https://doi.org/10.1093/nar/gkw1092>
36. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000; 28(1):27–30. <https://doi.org/10.1093/nar/28.1.27>
37. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res*. 2018; 46(D1):D661–D667. <https://doi.org/10.1093/nar/gkx1064> PMID: 29136241
38. UniProt Consortium T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. 2018; 46(5):2699. <https://doi.org/10.1093/nar/gky092>
39. Koscielny G, An P, Carvalho-Silva D, Cham JA, Fumis L, Gasparyan R, et al. Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res*. 2017; 45(D1):D985–D994. <https://doi.org/10.1093/nar/gkw1055> PMID: 27899665
40. Inzucchi S, Bergenstal R, Buse J, Diamant M, Ferrannini E, Nauck M, et al. Management of hyperglycemia in type 2 diabetes: a patient-centered approach. Position Statement of the American Diabetes

- Association (ADA) and the European Association for the Study of Diabetes (EASD). *Diabetes Care* 2012; 35:1364–1379. *Diabetes Care*. 2013;36(2):490–490. <https://doi.org/10.2337/dc12-0413> PMID: 22517736
41. Diez JJ, Iglesias P. The role of the novel adipocyte-derived hormone adiponectin in human disease. *Eur J Endocrinol*. 2003; 148(3):293–300. <https://doi.org/10.1530/eje.0.1480293>
  42. Yamauchi T, Kamon J, Waki H, Terauchi Y, Kubota N, Hara K, et al. The fat-derived hormone adiponectin reverses insulin resistance associated with both lipoatrophy and obesity. *Nat Med*. 2001; 7(8):941–946. <https://doi.org/10.1038/90984> PMID: 11479627
  43. Fischer S, Hanefeld M, Haffner SM, Fusch C, Schwanebeck U, Kohler C, et al. Insulin-resistant patients with type 2 diabetes mellitus have higher serum leptin levels independently of body fat mass. *Acta Diabetol*. 2002; 39(3):105–110. <https://doi.org/10.1007/s005920200027> PMID: 12357293
  44. Cummings BP. Leptin therapy in type 2 diabetes. *Diabetes Obes Metab*. 2013; 15(7):607–612. <https://doi.org/10.1111/dom.12048>
  45. Guerre-Millo M, Gervois P, Raspé E, Madsen L, Poulain P, Derudas B, et al. Peroxisome Proliferator-activated Receptor  $\alpha$  Activators Improve Insulin Sensitivity and Reduce Adiposity. *Journal of Biological Chemistry*. 2000; 275(22):16638–16642. <https://doi.org/10.1074/jbc.275.22.16638> PMID: 10828060
  46. Andrulionyte L, Kuulasmaa T, Chiasson JL, Laakso M. Single Nucleotide Polymorphisms of the Peroxisome Proliferator-Activated Receptor- $\alpha$  Gene (PPARA) Influence the Conversion From Impaired Glucose Tolerance to Type 2 Diabetes. *Diabetes*. 2007; 56(4):1181–1186. <https://doi.org/10.2337/db06-1110>
  47. Lee MR, Yang HJ, Park KI, Ma JY. *Lycopus lucidus* Turcz. ex Benth. Attenuates free fatty acid-induced steatosis in HepG2 cells and non-alcoholic fatty liver disease in high-fat diet-induced obese mice. *Phytomedicine*. 2019; 55:14–22. <https://doi.org/10.1016/j.phymed.2018.07.008>
  48. Wang M, Ma LJ, Yang Y, Xiao Z, Wan JB. n-3 Polyunsaturated fatty acids for the management of alcoholic liver disease: A critical review. *Crit Rev Food Sci Nutr*. 2019; 59(sup1):S116–S129. <https://doi.org/10.1080/10408398.2018.1544542>
  49. Nguyen LT, Chen H, Zaky A, Pollock C, Saad S. SIRT1 overexpression attenuates offspring metabolic and liver disorders as a result of maternal high-fat feeding. *J Physiol (Lond)*. 2019; 597(2):467–480. <https://doi.org/10.1113/JP276957>
  50. Kharitonov A, Shiyanova TL, Koester A, Ford AM, Micanovic R, Galbreath EJ, et al. FGF-21 as a novel metabolic regulator. *J Clin Invest*. 2005; 115(6):1627–1635. <https://doi.org/10.1172/JCI23606> PMID: 15902306
  51. Vallon V, Platt KA, Cunard R, Schroth J, Whaley J, Thomson SC, et al. SGLT2 mediates glucose reabsorption in the early proximal tubule. *J Am Soc Nephrol*. 2011; 22(1):104–112. <https://doi.org/10.1681/ASN.2010030246> PMID: 20616166
  52. Ren G, Kim JY, Smas CM. Identification of RIFL, a novel adipocyte-enriched insulin target gene with a role in lipid metabolism. *Am J Physiol Endocrinol Metab*. 2012; 303(3):E334–351. <https://doi.org/10.1152/ajpendo.00084.2012>
  53. Yi P, Park JS, Melton DA. Betatrophin: a hormone that controls pancreatic  $\beta$  cell proliferation. *Cell*. 2013; 153(4):747–758 <https://doi.org/10.1016/j.cell.2013.04.008>
  54. Yi P, Park JS, Melton DA. Retraction Notice to: Betatrophin: A Hormone that Controls Pancreatic  $\beta$  Cell Proliferation. *Cell*. 2017; 168(1-2):326. <https://doi.org/10.1016/j.cell.2016.12.017>
  55. Ebert T, Kralisch S, Hoffmann A, Bachmann A, Lossner U, Kratzsch J, et al. Circulating angiopoietin-like protein 8 is independently associated with fasting plasma glucose and type 2 diabetes mellitus. *J Clin Endocrinol Metab*. 2014; 99(12):E2510–2517. <https://doi.org/10.1210/jc.2013-4349> PMID: 25325797
  56. Guo K, Yu H, Lu J, Bao Y, Chen H, Jia W. Decreased serum betatrophin levels correlate with improved fasting plasma glucose and insulin secretion capacity after Roux-en-Y gastric bypass in obese Chinese patients with type 2 diabetes: a 1-year follow-up. *Surg Obes Relat Dis*. 2016; 12(7):1343–1348. <https://doi.org/10.1016/j.soard.2016.01.024>
  57. Zhang R. Lipasin, a novel nutritionally-regulated liver-enriched factor that regulates serum triglyceride levels. *Biochem Biophys Res Commun*. 2012; 424(4):786–792. <https://doi.org/10.1016/j.bbrc.2012.07.038>
  58. Vatner DF, Goedeke L, Camporez JPG, Lyu K, Nasiri AR, Zhang D, et al. Angptl8 antisense oligonucleotide improves adipose lipid metabolism and prevents diet-induced NAFLD and hepatic insulin resistance in rodents. *Diabetologia*. 2018; 61(6):1435–1446. <https://doi.org/10.1007/s00125-018-4579-1> PMID: 29497783
  59. Creative Commons: Attribution 4.0 International (CC BY 4.0) license. Available from: <https://creativecommons.org/licenses/by/4.0/>. [Accessed 21 November 2019].

60. Lee MS, Han HJ, Han SY, Kim IY, Chae S, Lee CS, et al. Loss of the E3 ubiquitin ligase MKRN1 represses diet-induced metabolic syndrome through AMPK activation. *Nat Commun*. 2018; 9(1):3404. <https://doi.org/10.1038/s41467-018-05721-4> PMID: 30143610
61. Han H, Chae S, Hwang D, Song J. Attenuating MKRN1 E3 ligase-mediated AMPK $\downarrow$  suppression increases tolerance against metabolic stresses in mice. *Cell Stress*. 2018; 2(11):325–328. <https://doi.org/10.15698/cst2018.11.164>
62. Francini F, Schinella GR, Rios JL. Activation of AMPK by Medicinal Plants and Natural Products: Its Role in Type 2 Diabetes Mellitus. *Mini Rev Med Chem*. 2019; 19(11):880–901. <https://doi.org/10.2174/1389557519666181128120726>