Combining mass spectrometry and machine learning to discover bioactive peptides

Madsen, Christian T.; Refsgaard, Jan C.; Teufel, Felix G.; Kjærulff, Sonny K.; Wang, Zhe; Meng, Guangjun; Jessen, Carsten; Heljo, Petteri; Jiang, Qunfeng; Zhao, Xin; Wu, Bo; Zhou, Xueping; Tang, Yang; Jeppesen, Jacob F.; Kelstrup, Christian D.; Buckley, Stephen T.; Tullin, Søren; Nygaard-Jensen, Jan; Chen, Xiaoli; Zhang, Fang; Olsen, Jesper V.; Han, Dan; Grønborg, Mads; de Lichtenberg, Ulrik

Published in:
Nature Communications

DOI:
10.1038/s41467-022-34031-z

Publication date:
2022

Document version
Publisher's PDF; also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Madsen, C. T., Refsgaard, J. C., Teufel, F. G., Kjærulff, S. K., Wang, Z., Meng, G., Jessen, C., Heljo, P., Jiang, Q., Zhao, X., Wu, B., Zhou, X., Tang, Y., Jeppesen, J. F., Kelstrup, C. D., Buckley, S. T., Tullin, S., Nygaard-Jensen, J., Chen, X., ... de Lichtenberg, U. (2022). Combining mass spectrometry and machine learning to discover bioactive peptides. Nature Communications, 13, [6235]. https://doi.org/10.1038/s41467-022-34031-z
Combining mass spectrometry and machine learning to discover bioactive peptides

Peptides play important roles in regulating biological processes and form the basis of a multiplicity of therapeutic drugs. To date, only about 300 peptides in human have confirmed bioactivity, although tens of thousands have been reported in the literature. The majority of these are inactive degradation products of endogenous proteins and peptides, presenting a needle-in-a-haystack problem of identifying the most promising candidate peptides from large-scale peptidomics experiments to test for bioactivity. To address this challenge, we conducted a comprehensive analysis of the mammalian peptidome across seven tissues in four different mouse strains and used the data to train a machine learning model that predicts hundreds of peptide candidates based on patterns in the mass spectrometry data. We provide in silico validation examples and experimental confirmation of bioactivity for two peptides, demonstrating the utility of this resource for discovering lead peptides for further characterization and therapeutic development.

Peptides are known to be potent regulators of a diverse set of biological functions and form the basis of around 80 marketed drugs, with an additional 150 peptide-based candidates in clinical development. Examples of well characterized bioactive peptides include GLP-1, Islet amyloid polypeptide (amylin) and Peptide YY (PYY) which regulate glucose stimulated insulin secretion, glycogen deposition in muscle, and colonic mobility, respectively.

Most bioactive peptides are formed by specific enzymatic cleavages of protein precursors, exemplified by pro-glucagon which is processed, in a tissue-specific manner, into at least nine different peptides with divergent functionalities. Peptide activity may furthermore be modulated by post-translational modifications or extracellular peptidases. For most of the well described peptide hormones, the precursor travels via the canonical secretory pathway where it is cleaved by prohormone convertases, trimmed by carboxypeptidase E and in some cases amidated by the PAM enzyme (requiring a Glycine in the +1 position). However, not all bioactive peptides follow this path (Supplementary Fig. 1). For instance, the antimicrobial peptide Buforin is proteolytically generated from histone H2A, illustrating that peptides can be formed even from processed precursors.
intracellular proteins with a function far removed from that of the peptide. Other peptides originate from cleavage of longer peptides, as illustrated by the vasoconstrictive peptide Angiotensin-II which is formed in the blood from the inactive Angiotensin-I peptide by the Angiotensin-converting enzyme (ACE). ACE also cleaves the vasodilator peptide Bradykinin into an inactive form and is the target of hypertension drugs, due to its role in regulating these potent bioactive peptides. When studying biological samples and whole tissues, one should thus expect to observe multiple variants of the same peptides, only some of which are bioactive, as well as differences in abundance, processing, secretion, and degradation across tissues.

Advances in high-throughput omics technologies have over the past decades shed light on the many layers of cellular regulation, and led to the systematic mapping of the human genome, transcriptome, and proteome, as well as the regulatory roles of RNA, epigenetic modifications, and post-translational modifications. In comparison, discovering bioactive peptides and uncovering their mode of action remains a major challenge and to date only about 300 peptides have been characterized in human. The majority of these functionally validated peptides, which include neuropeptides, endocrine peptides, or antimicrobial peptides, have been discovered via years of focused experimentation, rather than by large-scale comprehensive mapping. The fact that peptides are only formed and released in certain tissues and cell types under certain conditions—which may be unknown or difficult to reproduce experimentally—adds to the complexity of identifying bioactive peptides.

In recent years, mass spectrometry (MS) based peptidomics analysis has emerged as a powerful and sensitive tool for systematically mapping the peptide space (the peptidome) present in tissues or biofluids, including known bioactive peptides, longer inactive precursor peptides, shorter inactive forms, as well as the thousands of inactive degradation products formed as a consequence of natural protein and peptide turnover. MS-based peptidomics analysis is thus well-suited for monitoring the abundances of already known peptides in tissues and cellular extracts but using it as a tool for discovering novel bioactive peptides presents a major needle-in-a-haystack challenge of separating the small set of real bioactive peptides from the vast background of observed degradation products and inactive precursors. The sum of unique peptide sequences reported from small- and large-scale peptidomics experiments during the past decade far exceeds the capacity for functional testing. Methods for identifying and prioritizing the most promising candidates peptides for experimental validation are therefore needed to advance the field of peptide discovery.

In this study, we conducted a large-scale peptidomics analysis of seven different organs in four different mouse strains. Realizing that the majority of the observed peptide sequences are likely inactive degradation products, we developed a computational algorithm based on machine learning that uses the structure of the peptide clusters observed for each protein to predict the most likely endogenously functional peptides. We show that the algorithm can indeed identify many of the known annotated bioactive peptides directly from our MS data and that the other high-scoring candidate peptides suggested by the algorithm display characteristics similar to those of the known peptides, such as originating from secreted protein precursors and having known cleavage motifs in their flanking regions, even though the algorithm does not use these features as input for its predictions. To further enhance our ability to identify potential bioactive peptides, we combine the output of our prediction framework with in-silico bioactivity prediction based on amino acid composition to identify a number of interesting new candidate peptides. Using in-vitro and in-vivo screening models of diabetes, we confirm potential bioactivity of several of these predicted peptides. For each tissue investigated, we provide a list of the highest scoring peptides that collectively represent a resource ripe for further exploration and functional characterization. We lastly demonstrate that additional known peptides can be identified by combining our prediction framework with an algorithm that assembles observed degradation fragments into full length peptides.

Results

Experimental setup

Our experimental and computational workflow is illustrated in Fig. 1a. We investigated seven metabolically active tissues consisting of liver, muscle, intestine (ileum), brain, pancreas, epididymal fat, and subcutaneous fat from four different mouse strains commonly used in diabetes research: leptin receptor-deficient mice C57bl/ks-Leprdb/Leprdn (DB) versus C57bl/ks-Leprdb/+ (WT), and C57bl/6j mice fed a low-fat diet (LF) or high-fat diet (HF). To ensure statistical robustness of peptide abundance estimation, we used 12 animal replicates in each group, leading to a total of 336 samples. All tissue samples were heat-stabilized before homogenization to minimize post-mortem degradation of the in-vivo peptide. We elected not to reduce/alkylate the samples as our downstream automated peptide synthesis platform did not cope well with the complexity of intra- or inter-molecular sulfide bridges (Supplementary Fig. 2a). Peptides were subsequently separated from co-purified proteins using a molecular weight cut-off spin-filter which offered scalable and consistent peptide enrichment over proteins without any significant size bias below the cutoff (Supplementary Fig. 2b,c). Peptide mixtures were analyzed by online nanoflow liquid chromatography tandem mass spectrometry using high-resolution higher-energy collisional dissociation fragmentation. Raw files were searched with a selected number of post-translational modifications using two different search engines which resulted in the high-confidence identification of 157,857 unique peptide sequences across all tissues, conditions, and replicates (Fig. 1b). The output was filtered and combined (Supplementary Data 1), and accuracy, score distribution and sequence coverage were comparable to other peptidomics dataset (Supplementary Fig. 3a–j). The data set provide comprehensive coverage of different tissues and conditions under which peptides may be produced and secreted. It is, to our knowledge, the largest single peptidomics study to date in terms of tissue coverage and number of samples/replicates (Fig. 1c).

Tissue, diet and genetic background manifests at the peptide level

Although many of the observed peptides are found in multiple tissues, each tissue contributes with a unique set of peptides, with ileum being the largest contributor of unique sequences (Fig. 1b). Non-linear dimensionality reduction with UMAP reproduced the tissue and experimental groups (Fig. 2a,b) and clustering showed that tissue type was more strongly manifested in the data than strain background, with the two fat tissues being most closely related (Fig. 2c). Interestingly, different sequence motifs were observed in the flanking regions of the detected peptides from different tissues, in line with the expectation that enzymatic processing varies by tissue (Supplementary Fig. 4). An additional test experiment verified that the tissue-specific peptide signatures do not simply reflect degradation of highly abundant proteins (Supplementary Fig. 5a,b and Supplementary Data 2).

Sensitivity and degradation in peptidomics data

To assess the sensitivity and coverage of our data, we compared it to a curated list of 294 known and annotated peptides extracted from different peptide databases (Supplementary Data 3), and found that 89 (30%) of these appear in our MS data, including their known post-translational modification pattern, with an additional 76 known peptides covered partially by shorter fragments. Our mammalian peptidome analysis of 7 tissues thus covers nearly half of the annotated mouse peptides, as full or partial matches, demonstrating the
sensitivity of the technology and the potential of this data set as a resource for discovery of new bioactive peptides. The challenge, however, is that the known peptides are set against a vast background of peptides representing degradation fragments rather than real endogenously produced peptides with biological functions.

Mapping observed peptide sequences onto their cognate protein backbone proved to be a powerful way to illustrate the structure of the data, as exemplified for Secretogranin-1 (CgB); a protein of the granin family known to be processed into bioactive peptides\(^{16}\), as also observed in other recently published peptidomics studies as well (Fig. 3a). We observe several of the known Secretogranin-1 peptides as well as clusters indicating the existence of other uncharacterized peptides. The peptide clusters observed tend to align perfectly with flanking dibasic motifs\(^{10}\) (KR, RK, KK, RR, etc.) consistent with Secretogranin-1 being canonically cleaved by prohormone convertases in islets, pancreas, intestine and brain where these enzymes are known to process many other secreted containing dibasic cleavage sites. Such degradation is not specific to the known annotated peptides as shown in Supplementary Figs. 6a and 7a. The real active variant of the peptide automatically from the data. In some cases, longer unprocessed variants are also seen. On top of this complexity for the known bioactive peptides, routine breakdown of all some of the real active variant of the peptide automatically from the data. In some cases, longer unprocessed variants are also seen. On top of this complexity for the known bioactive peptides, routine breakdown of all some of the real active variant of the peptide automatically from the data. In some cases, longer unprocessed variants are also seen. On top of this complexity for the known bioactive peptides, routine breakdown of all some of the real active variant of the peptide automatically from the data. In some cases, longer unprocessed variants are also seen. On top of this complexity for the known bioactive peptides, routine breakdown of all some of the real active variant of the peptide automatically from the data. In some cases, longer unprocessed variants are also seen. On top of this complexity for the known bioactive peptides, routine breakdown of all.
(Fig. 4a; intensity start as example), as well as additional features that express amino acid modifications observed by MS (C-terminal amidation and N-terminal acetylation) (see Supplementary Note 1). We then trained various machine learning methods by using the set of known annotated peptides as positive examples (Supplementary Data 3) and the remaining observed peptides as negative examples. Each peptide was represented by its feature vector and nested five-fold cross-validation was used to ensure that the performance metrics reported for each model is based only on predictions on the hold-out test sets unseen by the methods during training and parameter optimization (Fig. 4b, Supplementary Note 1). The more advanced, non-linear models (Random Forest and SVMs) did not confer any advantage over their simpler linear counterparts (Bayesian and simple logistic regression; Fig. 4c, Supplementary Fig. 7a), and no performance benefit was observed with elastic net regularization or SMOTE (Supplementary Fig. 7b, c), despite the highly imbalanced training and test sets (few positive examples and many negative examples). We therefore chose to base our final method, that we term Predicted Peptide Variant (PPV), on simple logistic regression which offers a high degree of explainability and allows easy quantification of the importance and directionality of the input features. We found that the largest contribution to the model predictions comes from the features that encode the likely start and stop of a peptide cluster (Fig. 4b) and that C-terminal amidation adds positively to the predictions, consistent with this modification being required for the bioactivity of many known peptides37, including Neuromedin-C for which we confirmed the importance of the amidation for the function of the peptide (Supplementary Fig. 8a–c).

We compared the PPV model to a simple null model (AUC 0.732) which uses the abundance of each peptide as its only predictor. The Null model finds only 5 known training peptides among the 300 most abundant peptides, showing that the known peptides cannot be identified simply from their abundance. In comparison, the PPV model (AUC 0.886) identifies 48 known training peptides among its top 300 predictions, corresponding to a 176-fold enrichment over randomly picking peptides from the mass spectrometry data.

The PPV model was used to score and rank all observed peptides in each tissue, resulting in the identification of both known bioactive peptides and uncharacterized candidate peptides. The list of high-scoring peptides for each tissue is available as a supplementary table (Supplementary Data 5) and the source code for the model is openly available from GitHub (https://github.com/jancr/ppv).

It should be noted that although most of the annotated peptides extracted from peptide databases have validated functions, the positive training set (Supplementary Data 3) also contains peptides that are known to exist but for which no bioactivity has yet been reported, such as the C-terminal peptides of NPY and PYY. The PPV method is therefore not a predictor of bioactivity per se but rather a method suited for picking peptides from the mass spectrometry data.

Fig. 2 | Diet and genetic background manifested at the peptide level. a Tissue specificity on entire dataset. Non-linear dimension reduction employing UMAP with imputed values on 1% quantile of lowest intensity peptides61. Average values from 12 animals for each strain type: wildtype (WT), diabetic (DB), low-fat diet (LF), and high-fat diet (HF). b Strain specificity using UMAP with only valid values. Each datapoint is one animal. c Pearson correlation plot based on all pairwise peptide intensities (abundance) within groups. Median intensity values are used within each n = 12 group from the MaxQuant dataset.
extracts more signal from the structure of the MS-data than Pepti-DeRanker does from the amino acid composition towards recovering the known, annotated—and predominantly bioactive—peptides.

**Model predictions and in-silico validation**

To further build confidence in the PPV results, we analysed the flanking regions of the high-scoring predicted peptides and found a clear tissue-specific enrichment of sequence motifs (Fig. 5a). The positive training peptides were removed prior to this analysis to ensure that any signals observed reflect the properties of the new predicted peptides without bias from the peptides that the model was trained on. For brain, pancreas and ileum, the flanking regions show marked enrichment of the di-basic cleavage motifs recognized by prohormone convertases which are known be active in these particular tissues (Supplementary Fig. 9). Different motifs were observed in fat tissues, indicating that other enzymes may be active here. The peptides predicted by the PPV model also more often originate from secreted proteins (Fig. 5b). This aligns with the fact that most, albeit not all, annotated peptides come from precursors which contain an N-terminal signal peptide. Since none of the input features to the PPV model encode knowledge of the amino acid sequence of the flanking regions around the predicted peptides, or the secretion status of the parent proteins, the enrichment of these properties independently supports the validity of our methodology and predictions.

When applied to tens of thousands of peptides observed in a tissue, the probabilistic PPV model only predicts a small subset of peptides with a high probability of being real peptides (Fig. 5c). The number of very high-scoring (>0.05; red) and high-scoring (>0.01; yellow) peptides were 93/610 for brain, 39/481 for intestine (Ileum), 64/455 for pancreas, 20/229 for liver, 18/306 for quadriceps muscle, 34/346 for epidydimal fat, and 42/462 for subcutaneous fat, respectively (Supplementary Data 5). Since the PPV method is not a perfect
other peptides observed. Many of the predicted peptides identify known bioactive neuropeptides among tens of thousands of highest scoring peptides. The data and predictions confirm the existence of previously reported peptides but also shows that VGF processing is considerably more complex than hitherto described. An another data as well when evaluating these lists which are purely based on signals extracted from the MS peptidomics data.

Table 1 illustrates the predictions for the 25 highest scoring peptides found in brain tissue demonstrating the PPV method’s ability to identify known bioactive neuropeptides among tens of thousands of other peptides observed. Many of the predicted peptides identified share the flanking sequence motifs characteristic of neuropeptides (recognized by pro-hormone convertases).

The PPV predictions are illustrated visually for the neurosecretory protein VGF (Fig. 5d) using red and yellow colour to highlight the highest scoring peptides. The data and predictions confirm the existence of previously reported peptides but also shows that VGF processing is considerably more complex than hitherto described. An example is the brain-specific and highly conserved, 46-amino acid peptide (position 375-420) which we decided to call “GGGE-46” (Fig. 5d). This peptide is among the high-scoring in brain (PPV score 0.017) and its sequence is perfectly flanked by di-basic cleavage motifs (RR and KR). Interestingly, a previously reported shorter form (375-407) scores well below our threshold, illustrating the discriminatory power of the PPV algorithm. The model also correctly identifies the right variant of NERP-1 (rank 10 in brain with score 0.273) which is involved in body fluid homeostasis (in both brain and pancreas independently) and confirms the existence of the uncharacterized “24-63” peptide (rank 49, score 0.084), while scoring other variants within those clusters low. Both peptides were only annotated recently and therefore not included as positive examples during training (in fact, they were wrongly labelled as negatives). Their identification among the top 50 highest scoring peptides in brain (out of 20,597 observed) thus serve as independent “in-silico validation” of the PPV model’s ability to generalize beyond its training data (no overfitting). The correct identification of NERP-1 furthermore demonstrates that the model does not simply pick the longest variant in a cluster, since several low scoring variants were found to extend beyond the NERP-1 sequence. Other high-scoring in-silico validation examples that were not part of the positive training set include Serpinin-RR, ‘joining peptide’ from POMC, EH24 from pro-Thyrotropin-releasing hormone and IP2 from pro-pro-glucagon (Supplementary Data 5).

Another way to visualize and explore the data is to plot the results of the two orthogonal machine learning approaches (PPV based on MS data and PeptideRanker based on sequence composition) against each other for the highest scoring 300 peptides (Fig. 5e) using different coloring for the known peptides and the uncharacterized and potentially bioactive peptides. This nicely illustrates that many of the peptides identified via their expression pattern in our large MS data set (by the PPV method) have an amino acid composition similar to that of the
Fig. 5 | PPV model validation and predictions. a Logo-plot of N- and C-terminal flanking regions surrounding high scoring PPV predicted peptides in brain and pancreas. Lysine (K) and Arginine (R) are marked in dark blue, and Glycine (G) in light blue. Calculated by Kullback-Leibler divergence between PPV predictions with a score above 0.01 and the 80% lowest scoring background PPV predictions. All known annotated peptides were removed. b Secretion status determined by membership on protein precursor level based on signal peptide annotation and signalP predictions. Mouse proteome (Prot.), the global peptide database (Pept.) data in this study, PPV predictions (amino acid sequence composition) for the top300 PPV predictions (mass spectrometry signal) plotted against each other. Known annotated peptides used in the PPV training marked in red, and previously reported VGF peptides highlighted (yellow: >0.01, red: >0.05). Previously reported VGF peptides are mapped onto the cognate backbone at the top based on sequence similarity to annotated rat peptides. Ac Acetylation, Am Amidation. c PPV score density distribution in the data. X-axis is the PPV-score (Log10), and Y-axis is number of peptides. d Neurosecretory protein VGF (Uniprot ID: Q0VGU4) and tissue specific peptide patterns. Peptides are colour coded according to Log10 abundance (blue) and high-scoring predicted PPV peptides highlighted (yellow: >0.01, red: >0.05). Previously reported VGF peptides are mapped onto the cognate backbone at the top based on sequence similarity to annotated rat peptides. Ac Acetylation, Am Amidation. e PeptideRanker predictions (amino acid sequence composition) for the top300 PPV predictions (mass spectrometry signal) plotted against each other. Known annotated peptides used in the PPV training marked in red, unknown (negative set) marked in blue. Source data are provided as a Source Data file.
| PPV | ID      | Gene  | Peptide                  | N-    | Peptide Sequence | Am. | C-   | Sec. | Tissue(s) |
|-----|---------|-------|--------------------------|-------|------------------|-----|------|------|-----------|
| 0.670 | P41539:58-68 | Tac1  | Substance P [T]          | RIAR  | RPKDPQFSGLM      | Yes | GKR  | Yes  | B, I, P   |
| 0.588 | P55099:82-91 | Tac3  | Neurokinin-B [T]         | PQKR  | DMHDFPGVM        | Yes | GKR  | Yes  | B         |
| 0.533 | P56942:131-143 | Pmch  | Neuropeptide-Glu-Ile [T] | GEKR  | EIGDEENSAKFPX    | Yes | GRR  | Yes  | B         |
| 0.532 | P41539:98-107 | Tac1  | Neurokinin-A [T]         | SRKR  | HDTSPGVM         | Yes | GKR  | Yes  | B, I, P   |
| 0.455 | P47212:133-61 | Gal   | Galanin [T]              | KEKR  | GNTLNSAVYGLGPHAIWNRFSFSGGLT | Yes | GKR  | Yes  | B, I, P, F |
| 0.436 | P01193:110-110 | Pnoc  | Joining peptide (mis-annotated) | AQRG | AEEAVWGSSFSFPEFR | Yes | GKR  | Yes  | B          |
| 0.383 | P32648:101-107 | Vip   | Intestinal peptide PHI-27 [T] | KNAR  | HADQVEFDYRILGQ SAKYLES1 | Yes | GKR  | Yes  | B, I, P    |
| 0.317 | P32648:125-152 | Vip   | Vasopressin intestinal peptide [T] | PIKR | HSDAVPSTHRRLKRQMNKVIYLNILN | Yes | GKR  | Yes  | P, B, I    |
| 0.295 | Q03517:566-610 | Scg2  | Substance P [T]          | VSRK  | IPVGLKNEDETPNRQXI PEMMNKLYELMNQGRQAEQREHLA | No  | KRM  | Yes  | B, P, I, Sc, Ep |
| 0.273 | Q00004:1285-1309 | Vgf    | NERP-1 (delayed annotation) | KVVR  | LEGSFLGGSEAGERLQLQGLAQEA | Yes | GRR  | Yes  | B, P, Sc    |
| 0.264 | P41539:12-107 | Tac1  | Neuropeptide K [T]       | MKGR  | DADSVKVQALWALKLYGKQISHKRKRTDSFPGVM | Yes | GKR  | Yes  | B, I       |
| 0.244 | P13562:22-31 | Ghrh  | Gonadotropin-1 [T]       | GCCS  | QMSESGLPG        | Yes | GKR  | Yes  | B          |
| 0.243 | Q3TC46:403-430 | Fatll | Substance P [T]          | DPAK  | MIALQREKMQMS KIQMQCQSTDPFLDD | Yes | FYQ  | No   | B         |
| 0.230 | O55241:196-96 | Orcnt | Orexin-B [T]             | LGKR  | RPQPSQGQSLQRELQKQNAKHAQILMT | Yes | GRR  | No   | B         |
| 0.212 | P57774:48-97 | Apy   | C-terminating peptide of NPY [T] | YGKR  | SSPETLISLILKESNeAQPTRTEDPSMW | No  | ---- | Yes  | B, I, P, Ep, Sc, Q |
| 0.203 | O88935:144-151 | Syn1  | Substance P [T]          | D1KV  | RQAESFDL         | Yes | NLV  | No   | B         |
| 0.199 | BZ0004:1131-1154 | Plxn2b | Substance P [T]          | MTEL  | RAEAPVGAERTICTLDETOLYCE | Yes | PPEV | No   | B         |
| 0.198 | Q9Q30A:110-119 | Calca | Calcitonin-GP peptide (short form) | GLLS  | RGSGVKQNFVPVMTQGEAF | Yes | GRR  | Yes  | S, B, Ep, Q, I |
| 0.190 | Q63261:125-50 | Trh   | N-terminal fragment (unannotated) | KSCE  | LLEAQEGQAGFVLPGHLEPKVQVRPE | No  | RRFL | Yes  | B         |
| 0.160 | A8DOA4:12-46 | Hbbt1 | Substance P [T]          | _M    | VIHTDAEKAAGMSLVRGLVRVMDKPEGSCGEALQILVYQVYDSF | Yes | GDS  | No   | B, Q, Sc    |
| 0.159 | Q70251:12-46 | Setlin | Substance P [T]          | _M    | GPDLLTPAG        | Yes | LQVL | No   | B, Q, L, P  |
| 0.157 | P57774:27-64 | Apy   | Substance P [T]          | LAEG  | YPSKPVNPQIEA PAEAKRNYSAHLHYINLTQRQY | Yes | GKR  | Yes  | E, B, Sc, I, P |
| 0.153 | P09240:121-44 | Cck   | Propeptide (short form) | GALLA | QQPRAPAPAGVFDPQDSQKIQDGASDPARGDDQ | No  | RAVL | Yes  | B, I, P    |
| 0.141 | A2W4W3:327-349 | Gabre | Preptin (short form)    | EKPP  | EPEPOPEEPPEPEPOPEPPEPEPE | Yes | PKEU | No   | B         |
| 0.141 | Q9D3D9:23-46 | Atp5a | Substance P [T]          | ARTY  | ADEAAAAPAPAGKQMSFGSAFAAP | Yes | QVFF | No   | B         |

Top 25 highest scoring PPV peptides in brain. Uniprot protein identifier is separated from peptide position with a colon. The peptide column contains annotation status or a suggested name (SN). If followed by a [T] the peptide was included as positive training example. Other named peptide was not used for training but later found to match a previously described peptide (either due to delayed annotation or mis-annotation). Dibasic cleavage motifs are indicated in the N- and C-terminal flanking regions, with potential Glycines in position +1. C-terminal Amidation (Am.: Yes if >98% and No if <2%). Secretion status is indicated with Yes/No and tissue abbreviations are as follows; Brain (B), Ileum (I), Pancreas (P), Liver (L), EpiFat (Ep), ScFat (Sc), and Quadriceps muscle (Q).
known bioactive peptides (high PeptideRanker scores), which increases the likelihood that these are in fact bioactive.

Examples of uncharacterized peptides identified

Much to our surprise, the highest-scoring peptide in pancreas tissue (Supplementary Data 5; Pancreas) is a hitherto undescribed peptide found exclusively in pancreatic tissue where it is observed in all 48 mice (4 conditions x 12 replicates). The 22-amino acid peptide maps to an unreviewed protein precursor (Uniprot: D3Z596) originating from the insulin-2 gene. Its N-terminal sequence is identical to that of the connecting peptide (C-peptide) of insulin (including the known cleavage motif) but the rest of the sequence is distinctly different, and the peptide is amidated at its C-terminal. The sequence matches neither the normal insulin-2 precursor (P01326) nor that of insulin-1 (P01325), and we therefore considered if alternative splicing could explain our findings.

A splice site prediction analysis of the second intron of Insulin-2 suggests several possible acceptor sites (Fig. 6a), however only through splicing to a cryptic A2 acceptor site would the correct reading frame be established (Fig. 6b). This particular reading frame includes a C-terminal Glycine right before the stop-codon which is likely converted by the PAM enzyme to the C-terminal amidation we observe in our data (Supplementary Fig. 10a–c). This suggests a possible secretory role for the alternatively spliced peptide, that we termed Disjoining-peptide (D-peptide). Curiously, the D-peptide is scored 36 times higher than the classical C-peptide by the PPV algorithm, indicating that the D-peptide, not the C-peptide, is the one which stands out most clearly in the MS-data. The D-peptide is also scored relatively high by PeptideRanker (0.495) in support of a potential bioactive role for the peptide.

Some of the peptides are found independently in multiple tissues and we therefore combined the prediction probabilities from

Fig. 6 | A unique high-scoring peptide formed through alternative splicing in insulin. a Insulin-2 (Uniprot: P01326; NCBI sequence: NC_000073.7: c142233463-142232393) from Mus musculus strain C57BL/6j. Splicing and pro-peptide processing will generate the mature insulin’s A- and B-chain and the connecting peptide (C-peptide). The splicing in intron2 takes place between donor site 1 (D1) and acceptor site 4 (A4) as shown on the right. Introns are in small letters, exons is in capital letters. The ASSP and NetGene2 splice site predictions show the relative strength of donor sites (underlined; blue) and acceptor sites (boxed; red) normalized to D1.

b The Ins2-202 transcript (Ensembl ID: ENSMUST0000105930.8) match the Uniprot entry D3Z696, and is a 79 amino acid variant of insulin. Alternative splicing between the D1 donor site and the cryptic A2 acceptor site introduces a new reading frame with a premature stop-codon forming a pro-peptide consisting of an intact B-chain and a uncharacterized peptide (D-peptide) after prohormone convertase cleavage. The PAM enzyme converts the terminal Glycine in 2 sequential steps to a C-terminal amidation. Bottom right contains an alignment between the classic C-peptide and the D-peptide.

A splice site prediction analysis of the second intron of Insulin-2 suggests several possible acceptor sites (Fig. 6a), however only through splicing to a cryptic A2 acceptor site would the correct reading frame be established (Fig. 6b). This particular reading frame includes a C-terminal Glycine right before the stop-codon which is likely converted by the PAM enzyme to the C-terminal amidation we observe in our data (Supplementary Fig. 10a–c). This suggests a possible secretory role for the alternatively spliced peptide, that we termed Disjoining-peptide (D-peptide). Curiously, the D-peptide is scored 36 times higher than the classical C-peptide by the PPV algorithm, indicating that the D-peptide, not the C-peptide, is the one which stands out most clearly in the MS-data. The D-peptide is also scored relatively high by PeptideRanker (0.495) in support of a potential bioactive role for the peptide.

Some of the peptides are found independently in multiple tissues and we therefore combined the prediction probabilities from
individual tissues into a consolidated score for each peptide across all tissues, as an alternative way of viewing the prediction output (Supplementary Data S; PPV combined). This approach highlighted an N-terminal peptide derived from the glycolytic enzyme Tpi1 that the PPV method identifies consistently across brain, pancreas, muscle and fat tissues. The peptide, that we named APTR-15, contains the substrate binding residues of Tpi1 and is structurally followed by an exposed loop containing a prohormone convertase GRK motif consistent with its observed C-terminal amidation (Supplementary Fig. 11a–c). It should be noted that the enzyme is intracellular and would not under normal circumstances be co-localized with the prohormone convertases in the secretory pathway. Our data nonetheless strongly indicate that this particular peptide is formed in-vivo across multiple tissues. A potential bioactive role for this peptide is supported by PeptideRanker which scores it higher (0.844) than well described peptides like pancreatic hormone, Orexin-B, Gastrin-releasing peptide and VIP (Fig. 5e, Supplementary Data S).

Non-Glycine mediated amidation

Some of the high scoring peptides are found (in our MS data) to be amidated at the C-terminus without the Glycine (in position +1) that would normally be recognized and converted by the PAM enzyme (Table 1, Supplementary Data S). Such PAM-independent peptide amidation has been reported in several other studies25–27 and detailed inspection of the MS/MS ion spectra of selected examples from our study indeed support C-terminal amidation (Supplementary Fig. 12a–h). We cannot completely exclude the possibility of side-chain Glutamic acid to Glutamine conversion among a smaller subset of Glutamic acid containing peptides where the fragment ion series does not unambiguously confirm the amidation to be positioned in the C-terminal. A closer look at the sequence context around these non-Glycine amidations (Fig. 7a) reveal a decrease in dibasic amino acids, indicating that they are likely not products of prohormone convertases in the canonical secretory pathway. These peptides are instead enriched in a E/P motif around the site of amidation (Fig. 7a,b). These findings apply across the global data set as well as among those peptides selected by our algorithm. Although the enzymatic or non-enzymatic amidation mechanism remains to be elucidated, the data indicates that the majority of these peptides cannot be explained as random artifacts of the experimental procedures or the algorithm.

Experimental validation of predicted peptides in diabetes-related models

The computational framework developed and presented in this study can be used to shortlist the most promising peptides of which many should statistically be expected to be bioactive (Supplementary Data S,6). Validating such bioactivity and uncovering the mode of action of a candidate bioactive peptide is, however, inherently difficult, as it requires some initial functional hints to guide the choice of assays or models for confirmation. Such hints could be obtained from specialized sequence-based prediction tools like MultiPep40 although they still predict fairly broad functional categories (e.g. “neuropeptide”) that may have limited utility for guiding assay selection.

To authenticate the bioactivity of some of our predicted peptides, we relied on a set of diabetes-related assays and models already established in house and validated using known metabolic peptide hormones (Supplementary Fig. 13a–d) and selected non-functional peptides (Supplementary Fig. 14a–d). We therefore sought to increase the chance of a functional read-out in these particular assays by selecting candidates that originate from pre-cursor proteins/genes with functional links to diabetes, obesity and metabolism in public databases such as Gene Ontology (geneontology.org)41, DISEASES (diseases.jensenlab.org)42 and PHAROS (pharos.nih.gov)43. This approach obviously has limitations in cases where such information is scarce or where the function of the peptide is far removed from that of the parent protein/gene-level to which such data is linked (Supplementary Fig. 1).

Selected peptides were chemically synthesized or expressed in E. coli and tested for activity in vitro, as well as injected subcutaneously in db/db mice to quantify acute changes in relative blood glucose (BG) in vivo. One of the predicted peptides that displayed potential bioactivity was a previously undescribed peptide from Secretogranin-1 that we termed NHPD-50 (position 386-435). It has a high PPV score (0.028) and is flanked by dibasic KR-motifs (Fig. 3a, Supplementary Data S). Secretogranin-1 (CgB) is a component of the neuroendocrine secretory granules, and peptides from the granin family are known to regulate catecholamine, islet amyloid polypeptide and insulin release44. The NHPD-50 peptide reduced BG significantly four hours after single dose administration in db/db mice (Fig. 8a) and a significant increase of insulin in plasma was observed six hours after administration (Fig. 8b). The finding was replicated in a follow-up experiment using twice daily dosing (Fig. 8c) which showed a sustained BG lowering effect of this peptide over 8 h. In contrast, injection of lower scoring peptides did not elicit any response in these in-vivo models (Supplementary Fig 14a–d).
We did not observe a glucose stimulated insulin secretion (GSIS) response in INS1E cells or in mouse primary islets suggesting that the effect on blood glucose and the peptide-mediated plasma insulin increase acts through an indirect mechanism. To explore if an engineered NHPD-50 peptide variant with longer half-life could induce sustained BG reduction, we generated an N- and C-terminal albumin protracted version and expressed the construct in db/db mice using hydrodynamic gene delivery (Supplementary Fig. 15a). The protracted peptide did not confer sustained BG lowering effect in this experiment (Supplementary Fig. 15b,c). We did, however, observe a significant reduction of 3-hydroxybutyric acid in plasma suggesting a potential role in fatty acid metabolism (Supplementary Fig. 15d). In summary, our experiments support that the predicted NHPD-50 peptide displays bioactivity, although the data is not sufficient to decipher the exact biological mechanism of the new peptide.

Despite attempts to discover the potential biological function of the alternatively spliced insulin D-peptide, we did not observe any acute changes in BG, suggesting that the D-peptide probably maintains other functions unrelated to insulin or the classical C-peptide.

**Fragment assembly identifies known and uncharacterized peptides**

For some of the known bioactive peptides without a perfect full-length match, we observed shorter peptides that collectively covered the sequence. We therefore devised an algorithm which assembles overlapping fragments into in-silico predicted full-length peptides and found an additional 17 matches to known peptides including Amylin, Glicentin, Gastrin-releasing peptide, Somatostatin-28, PACAP-related-peptide and the CART peptide, demonstrating the potential of fragment assembly for identifying real bioactive peptides.

We therefore combined the assembly algorithm with the PPV model into a method termed PPV-assembly (Supplementary Note 1) with the aim of discovering additional peptides from our data (Supplementary Fig. 16a, b). The assembly creates on average 5 extra synthetic peptides for every observed peptide (expanding the data set from ca. 150,000 sequences to 1.2 million in total), many of which are likely not real (Supplementary Data 6) and we therefore found the PPV-assembly approach most useful in cases where other sources of information points to a particular protein of interest.

As an example, the model combined fragments in-silico to predict a new peptide in Gelsolin (Supplementary Fig. 17). Gelsolin is a calcium-regulated protein secreted into the blood facilitating pancreatic β-cell proliferation and insulin secretion through actin remodelling and syntaxin-4 interaction. Its N-terminal domain (28–161) has been suggested to normalize BG in diabetic mice. When administering this PPV-assembled peptide (position 567–625) to 3T3L1-MBX cells, we observed a dose-dependent increase in glucose uptake (Fig. 8d) and a dose-dependent inhibition of PEPCK expression in primary hepatocytes. N = 30,000 cells/well in two independent biological experiments. Cells treated for 4 h with serial diluted Gelsolin peptide (567–625) or serial diluted Insulin. Pck1 expression was normalized to Rplp0. Source data are provided as a Source Data file.
Discussion

The complex task of discovering new bioactive peptides is a two-step process, as one needs first to identify likely bioactive candidates and secondly employ the correct assay to unequivocally authenticate bioactivity.

In this study, we provide a resource to support the first step by sharing lists of candidate peptides (observed and assembled) predicted to be real and potentially bioactive by our machine learning framework (the PPV method) which we apply to a large whole-tissue peptidomics study of seven tissues that—to our knowledge—is by far the largest study conducted to date in terms of tissues, replicates and samples. The benefits of this approach are described above but here we also wish to highlight the limitations and biases of the resource we share with the broader scientific community.

Firstly, many important tissues and biological conditions are not covered in our study, meaning that certain peptides may be missed entirely. Secondly, we initially tested the use of reduction and alkylation but found that this only provided marginal improvements in the identification of cysteine-containing peptides at the expense of poorer identification rates of other peptides. Taking into account the need to disentangle intramolecular cysteine-bridges and the difficulty in chemically synthesizing such peptides, we chose to try to reduce and alkylate. Some known cysteine-containing peptides, like insulin, are thus absent from our analysis. The same applies to peptides with certain post-translational modifications, like the active form of CCKS which is sulfated on Tyrosine. To avoid a combinatorial explosion in the search space, the MS analysis was performed with only Oxidation (M), acetylation (N-term), amidation (C-term) and pyroglu (Q/E) as variable modifications. The PPV method does identify the correct unmodified variant of CCKS illustrating both the strength and limitation of the approach, as well as the difficulty in authenticating bioactivity which depends on testing the correct bioactive form in the right assay.

The machine learning based algorithms (PPV, PeptideRanker, etc.) are also not perfect, and should be expected to produce both false negatives (bioactive peptides missed) and false positives (non-functional peptide variants and degradation products predicted to be real and/or bioactive). The chance of the latter is arguably lower among peptides originating from secreted proteins and/or those precursors already known to form other bioactive peptides. We therefore provide the readers with the option to filter on these properties in the lists of candidate peptides (Supplementary Data 5,6). Vice versa, more false positives should be expected among predicted peptides originating from cytosolic proteins and we also note that the PPV model may have a slight bias towards peptides that start right after an initiator Met/homine. Nonetheless, we found it relevant to share all predictions, since some bioactive peptides are known to be produced from intracellular proteins (Supplementary Fig. 1).

Another limitation of our approach is that both machine-learning frameworks used in this study (PPV and PeptideRanker) are trained to identify bioactive peptides as a general class, and hence cannot be used to pinpoint the exact biological or molecular function of a peptide to guide assay selection. We did not in this work explore prediction of more fine grained functional categories owing to the relatively small set of positive examples (bioactive peptides) available to train the machine learning framework on.

In this work, we used other contextual information to select candidate peptides to test in a set of assays available to us. A larger systematic study would be required to assess the validity of this approach as a generalized concept for matching candidate peptides and functional assays.

The experimental validation of bioactivity also comes with its own set of limitations. Even in cases where the right bioactive form of a peptide is identified and tested in a suitable model, the functional readout may still be absent due to formulation issues, short peptide half-life, or the absence of other complementary factors required for full activity.

The central idea behind our PPV method is to turn the degradation “noise” that characterizes MS-based peptidomics studies into an asset rather than a challenge by learning from the global and local structure of the data. The concept could potentially be improved with new or improved feature encoding, unsupervised representation learning, and more refined approaches for assembling peptides in-silico.

In conclusion, our peptide data and predictions provide a comprehensive resource amenable for functional exploration beyond diabetes and suggest that more peptides than previously recognized could be functional entities with potential regulatory roles.

Methods

Animal experiments

Animal studies were carried out in accordance with the Danish Act on Experiments on Animals - Appendix A of ETS 123 and EU Directive 2010/63. Animal protocols was approved by the Institutional Animal Care and Use Committee of Novo Nordisk Research Center China and Ethical Review Council in Novo Nordisk A/S (Permission no. 2015-15-0201-00616). Animals were purchased either from Taconic (lean C57bl/6J mice), Jackson Laboratory (diet induced obese C57bl/6J mice fed 60% high fat diet (HF) and C57bl/6J mice fed 10% low fat diet (LF) or Charles River (C57bl/Ks db/db and C57bl/Ks db/db/+). Mice were all male and 12 weeks of age upon arrival and housed under standard conditions including a 12 h–12 h light–dark cycle, ~21 °C, and water and food ad libitum. All mice were anaesthetized using isoflorulane/O2/NO2 and following tissue extraction euthanized by decapitation at 15 weeks (C57bl/Ks db/db “DB” and C57bl/Ks db/db/+ “WT”) or at age 26 weeks (“HF” and “LF” mice). Collected organs (subcutaneous fat, epidydimal fat, pancreas, brain, liver, quadriceps muscle and gut ilium) were quickly rinsed in ice cold PBS and freeze clamped with Nitrogen-cooled iron. For the perfusion experiment anesthetized animals were transcardially perfused with an isotonic saline solution containing protease inhibitors. Other organs were fast frozen in 2-methyl-butane and stored at ~80 °C. Male diabetic BKS.Cg-Dock7m/+ Leprdb/+ (stock no: 000642) strain introduced from the Jackson Laboratory USA was used for blood glucose measurements. Upon arrival, mice were housed as 5 mice per cage under a condition of 12 h light/dark cycle with controlled temperature (26 ± 2 °C), and water and food ad libitum. The machine learning based algorithms (PPV, PeptideRanker, etc.) are also not perfect, and should be expected to produce both false negatives (bioactive peptides missed) and false positives (non-functional peptide variants and degradation products predicted to be real and/or bioactive). The chance of the latter is arguably lower among peptides originating from secreted proteins and/or those precursors already known to form other bioactive peptides. We therefore provide the readers with the option to filter on these properties in the lists of candidate peptides (Supplementary Data 5,6). Vice versa, more false positives should be expected among predicted peptides originating from cytosolic proteins and we also note that the PPV model may have a slight bias towards peptides that start right after an initiator Met/homine. Nonetheless, we found it relevant to share all predictions, since some bioactive peptides are known to be produced from intracellular proteins (Supplementary Fig. 1).

Another limitation of our approach is that both machine-learning frameworks used in this study (PPV and PeptideRanker) are trained to identify bioactive peptides as a general class, and hence cannot be used to pinpoint the exact biological or molecular function of a peptide to guide assay selection. We did not in this work explore prediction of more fine grained functional categories owing to the relatively small set of positive examples (bioactive peptides) available to train the machine learning framework on.

In this work, we used other contextual information to select candidate peptides to test in a set of assays available to us. A larger systematic study would be required to assess the validity of this approach as a generalized concept for matching candidate peptides and functional assays.

The experimental validation of bioactivity also comes with its own set of limitations. Even in cases where the right bioactive form of a peptide is identified and tested in a suitable model, the functional readout may still be absent due to formulation issues, short peptide half-life, or the absence of other complementary factors required for full activity.

The central idea behind our PPV method is to turn the degradation “noise” that characterizes MS-based peptidomics studies into an asset rather than a challenge by learning from the global and local structure of the data. The concept could potentially be improved with new or improved feature encoding, unsupervised representation learning, and more refined approaches for assembling peptides in-silico.

In conclusion, our peptide data and predictions provide a comprehensive resource amenable for functional exploration beyond diabetes and suggest that more peptides than previously recognized could be functional entities with potential regulatory roles.

Methods

Animal experiments

Animal studies were carried out in accordance with the Danish Act on Experiments on Animals - Appendix A of ETS 123 and EU Directive 2010/63. Animal protocols was approved by the Institutional Animal Care and Use Committee of Novo Nordisk Research Center China and Ethical Review Council in Novo Nordisk A/S (Permission no. 2015-15-0201-00616). Animals were purchased either from Taconic (lean C57bl/6J mice), Jackson Laboratory (diet induced obese C57bl/6J mice fed 60% high fat diet (HF) and C57bl/6J mice fed 10% low fat diet (LF) or Charles River (C57bl/Ks db/db and C57bl/Ks db/db/+). Mice were all male and 12 weeks of age upon arrival and housed under standard conditions including a 12 h–12 h light–dark cycle, ~21 °C, and water and food ad libitum. All mice were anaesthetized using isoflorulane/O2/NO2 and following tissue extraction euthanized by decapitation at 15 weeks (C57bl/Ks db/db “DB” and C57bl/Ks db/db/+ “WT”) or at age 26 weeks (“HF” and “LF” mice). Collected organs (subcutaneous fat, epidydimal fat, pancreas, brain, liver, quadriceps muscle and gut ilium) were quickly rinsed in ice cold PBS and freeze clamped with Nitrogen-cooled iron. For the perfusion experiment anesthetized animals were transcardially perfused with an isotonic saline solution containing protease inhibitors. Other organs were fast frozen in 2-methyl-butane and stored at ~80 °C. Male diabetic BKS.Cg-Dock7m/+ Leprdb/+ (stock no: 000642) strain introduced from the Jackson Laboratory USA was used for blood glucose measurements. Upon arrival, mice were housed as 5 mice per cage under a condition of 12 h light/dark cycle with controlled temperature (26 ± 2 °C), and water and food ad libitum.
Shotgun proteomics
For brain proteomics experiments n = 3 of each strain were used. Proteins were extracted with 6 µL/mg of 6 M Gnd-HCl. 5 mM TCEP, 10 mM CAA, 10 mM Tris-Cl pH = 8.0, 1 mM EDTA lysis buffer after denaturing heat stabilization of the brain sample. One mg protein from each sample was reduced/alkylated and subsequently digested with 5 µg Lys-C for 2 h and 5 µg Trypsin in 1% SDC, 100 mM Tris-Cl pH 8.0 overnight at 37 °C using a FASP approach with a 10 kDa MWCO spin filter. The tryptic peptides were subsequently cleaned up on a C18 SepPak column using standard methods. 50 µg of each digest was subsequently fractionated into 20 fractions using a Waters acuity CSIL 1.7 µm 1.0 × 15 mm column on a Ultimate 3000 HPLC ( Dionex, Sunnyvale, CA, USA) operating at 30 µL/min. Buffer A consisted of 5 mM ammonium bicarbonate and buffer B consisted of 5% ammonium bicarbonate, 90% acetonitrile. Proteomics data was searched using MaxQuant (ver. 1.5.6.2).

Mass spectrometry analysis
Peptides were eluted into 96-well microtiter plates with 20 µL 40% MeCN/0.5% acetic acid followed by 20 µL 60% MeCN/0.5% acetic acid. Peptides were reconstituted in 10 µL 2% MeCN, 0.5% acetic acid, 0.1% TFA after vacuum centrifugation in a speed-vac. Five microliters of the peptide eluate was separated by a linear MeCN gradient for 160 min in a 15-cm fused-silica emitter packed with reversed-phase ReproSil-Pur CIS-AQ 1.9 µm resin (Dr. Maisch GmbH) using a nanoflow Easy-nLC system (Thermo Scientific). The LC was connected through a nano-electrospray ion source to the mass spectrometer. We either used a QExactive orbitrap or QExactive HF orbitrap instrument using a top6 precursors setting. Calibration curves were made with Seaborn (ver. 0.11.2) in Jupyter (ver. 6.4.7). The PPV source code can be downloaded freely from: https://github.com/jancr/ppv. Supplementary Data 5, 6 contains all PPV predictions and PPV-assembly predictions with a score higher than 0.01.

Bacterial peptide production
DNA sequences encoding peptides above 35 amino acids in length were fused C-terminally to a 6×His-SUMO-tag and cloned into pET11d-derived plasmid. The plasmid was transformed into an E. coli BL21(DE3)-derived host strain, HNC54, expressing the NucB nuclease from Staphylococcus aureus (UniProtKB: A0A447ZB30). To over-express 6×His-SUMO-fusions, individual clones were inoculated in a 96-well format containing 1.2 mL Terrific Broth containing ampicillin in each well and cultivated at 37 °C with shaking at 800 rpm for 4.5 h. IPTG was added to 0.5 mM final concentration to induce expression and cultivation continued at 37 °C for 16 h. Cells were harvested by centrifugation and resuspended in 100 µL PBS pH 7.4 and freeze-thawed twice. Resulting cell lysate was mixed with 1.2 mL PBS pH 8.0 containing 8 M urea and 10 mM imidazole and centrifuged and the supernatant loaded on a Ni-NTA resin (QIAGEN) column pre-equilibrated with PBS pH 8.0 containing 10 mM imidazol. After washing with 20 bed volumes of PBS buffer, the fusion peptide was eluted with PBS pH 8.0 containing 300 mM imidazole. A PD10 column (GE Healthcare Life Sciences) was used to change the buffer into PBS containing 20 mM imidazole pH 8.0. The eluted fusion-peptide was digested with a SUMO protease at 1:500 (w/w, enzyme/protein) at room temperature overnight. The peptide was recovered in the flow-through fraction using a Ni-NTA column, and purified with Source 30 reverse-phase chromatography (GE Healthcare Life Sciences). Briefly, the Source 30 RPC column was washed with 5 bed volumes of 20 mM ammonium bicarbonate, and washed with 10 bed volumes of 0.5 M arginine at pH 9.0. The peptide was eluted with 40% isopropanol in 20 mM ammonium bicarbonate and the eluate evaporated with SpeedVac (ThermoFisher Scientific) to less than 0.5 mL, and then lyophilized to dryness. The lyophilized product was reconstituted in PBS pH 7.4. Peptides used for in-vivo dosing were checked with respect to purity, molecular mass, and endotoxin levels (<1 EU/kg animal)
using RP-HPLC and Kinetic Turbidimetric LAL testing according to the 
supplier (Charles River Laboratories).

**Peptide array synthesis**

Synthesis were carried out in four 96-well plates using an Intavis 
Multipep RSi synthesizer using Fmoc based solid phase peptide 
synthesis (SPPS). For C-terminal amide Fmoc-PAL AM resin from 
Novabiochem was used and for C-terminal acids the corresponding 
preloaded TGT resins from Novabiochem were used. Each peptide was 
synthesized on 5 µmol scale based on resin loading. The synthesis 
protocol used double deprotection using 20% piperidine + 0.1M 
oxyma in DMF and triple couplings using 6 equivalents amino acid/ 
DIC/oxyma 1:1:1 in each coupling and for a total of 3 h coupling 
time. Cleavage of peptides from resin was carried out using TFA/DTT/TIPS/ 
water 96:2:2:2 (1.2 mL) and the cleaved peptides were collected in 
corresponding 96-well deep well plates. The volume of TFA was 
reduced under a stream of nitrogen to 200-300 µL per well. The pep-
tides were precipitated with 1.2 mL ether and filtered and washed in 
deep-well Solvinert plates with a hydrophilic membrane. The crude 
peptides were dried using vacuum and redissolved in 500 µL DMSO 
and collected in 96-well plates. All peptides were analysed by UPLC-MS 
for assessment of purity. The DMSO stocks were lyophilized, then 
redissolved in acetonitrile/water 1:1 (1 mL) and then lyophilized again 
to produce the peptides as white powders.

**Formulation of peptides**

Freeze-dried peptides were rehydrated and formulated on 96-well 
plates. Endotoxin contents of 2 formulations per well plate were 
measured using LAL assay to rule out bacterial contamination. Rehy-
dration buffer for each peptide was based on its calculated isoelectric 
point (pI) so that its pH was ≥1 pH unit away from pl. Peptide con-
centration and purity was checked after rehydration with RP-UPLC 
using Acquity BEH C18 1.7 µm 2.1 × 30 mm column (Waters) and 
detection at UV215 nm. Mobile phases A & B consisted of 0.1% TFA in 
H2O and 0.1% TFA in 80% ACN/H2O, respectively, and elution was 
carried out from 20% to 95% phase A at 30 °C. Peptides with <40% 
purity were discarded, and the rest formulated at 300 µM in 50 mM 
phosphate at pH 6.0, 7.0 or 8.0, 12 mg/mL propylene glycol and 
0.5 mg/mL polysorbate 20 and sterile 
phosphate at pH 6.0, 7.0 or 8.0, 12 mg/mL propylene glycol and 
redissolved in acetonitrile/water 1:1 (1 mL) and then lyophilized again 
to produce the peptides as white powders.

**Peptide administration and blood glucose measurement**

Male diabetic BKS.Cg-Dock7m+Leprdb/J (stock no: 000642, Jackson 
Laboratory, USA) older than 11 weeks and with blood glucose levels 
>16 mM were selected and allocated to different treatment 
groups by randomization based on blood glucose. The mice body 
fractionation mice were anesthetized with isoflurane and blood was taken 
from the orbital puncture and transferred into two EDTA coated tubes 
and euthanized.

**Surgery and liver perfusion**

Perfusion buffers were pre-warmed to 37 °C and gassed with 95:5 
O2/CO2, for approximately 15–30 min prior to surgery. 8 weeks old male 
Sprague Dawley CrlCD(SD) rats (Charles River) were anesthetized with 
a mixture of N2/O2/isoflurane. Once animals had reached a surgical 
plane of anaesthesia, a U-shaped incision was made from the rib cage 
to the lower abdomen through the skin and muscle layer leaving the 
ribcage and diaphragm intact. The viscera were exposed, and the 
stomach and intestines were displaced to the right to reveal underlying 
blood vessels. The hepatic portal vein was isolated and cannulated with 
a 21G luer adapter connected to prefiltered catheter immersed in 
250 mL of perfusion buffer 1 (Kreb’s Ringer with 20 mM glucose, 
120 mM NaCl, 2.8 mM NaHCO3, 20 mM glucose, 1% Soll C, 5 mM 
HEPES, 100 µg/mL EGTA, and pH adjusted to 7.4). Perfusion buffer 1 was 
slowly infused at a rate of 5 mL/min using a peristaltic pump. The 
diaphragm was punctured to expose the thoracic cavity and the aorta 
severed. The flow rate was then slowly and steadily increased to 25 mL/ 
min. A vacuum pump was used to remove perfusion/blood overflow. 
Perfusion was continued until less than 5 mL of the initial 250 mL 
volume of perfusion buffer remains. Perfusion was then switched to a 
perfusion buffer 2 (120 mM NaCl, 2.8 mM NaHCO3, 20 mM glucose, 1% 
Soll C, 5 mM HEPES, 1.4 mM CaCl2, adjusted to pH 7.4) with a 
collagenase blend (7 mg/mL, Liberase TM Research Grade, Roche). Per-
fusion was continued until less than 5 mL of the 250 mL perfusion 
buffer remains. The liver was then immediately excised and processed 
for hepatocyte isolation. At the completion of the procedure the heart 
was removed to ensure euthanasia.

**Hepatocyte isolation**

After perfusion, the rat liver was excised and transferred to sterile dish 
containing 20 mL wash medium (Gibco #11043023). 5.5 mM glucose, 
supplemented with 5 mL Penicillin/Streptomycin 1% (Gibco #15140-
114), 100 nM dexamethasone, 1% FBS (Ausbion #VS500T), and 1 nM 
insulin (in house produced), and teared with forceps. Cells were dis-
persed by gently aspiration and filtered through a 100 µm nylon filter 
into a 50 mL tube. Cells were pelleted at 100 g at 4 °C for 3 min. 
Supernatant was decanted and cell pellets dispersed by aspirating 
gently with a bore pipet in 10 mL wash medium. Washing and resus-
pension steps was repeated for 3 times. Isolated hepatocytes were 
suspended in 20 mL Basal media (as wash medium but with 4% FBS). 
Cell number and viability were counted with a haemocytometer. 
Hepatocytes were diluted and plated into collagen-coated wells 
(30,000/well) and cultured in basal media for 4 h, then changed to 
starvation media (as wash medium with 0.1% FBS) overnight.

**PEPCk qPCR**

Primary hepatocytes were prepared as described above. Cells were 
treated with serial diluted peptide, 10 nM insulin and 10 nM glucagon 
(both controls in house produced) for 4 h. After treatment, cells were 
lysed, and RT-PCR was performed with Cells-to-CT Kit (Ambion 
#4402955) according to manufacturer’s recommendation. In brief, 
cells were lysed with 25 µL lysis buffer with shaking at 750 rpm for 
10 min before 2.5 µL stop solution was added. Reverse transcription 
was performed in 384 well-PCR plate. qPCR analysis was performed 
with Applied Biosystems Viia7 instrument (Life technologies). G6pc 
and Pck1 expression in H4IE cells were measured and normalized to 
Rplp0. PEPCk, G6PC and Rplp0 qPCR primers were purchased from 
ThermoFisher; PCK1 TaqMan Gene Expression Assays (Rn01529014 
mL), G6PC TaqMan Gene Expression Assays (Rn00689876 m1), RPLP0 
TaqMan Gene Expression Assays (Rn03302271 gH).
Glucose uptake assay
3T3-L1 MBX fibroblasts (ATCC, #CRL3242) were seeded at 20,000 cells per well in collagen coated 96-plates (Corning #356650) in DMEM (Gibco #10569) with 10% FBS (Ausbio #V5SS0T). Cells were differentiated into adipocytes for 4 days with 3% FBS, 1 µg/mL human insulin (in-house produced), 0.5 mM isobutylmethylxanthine (Sigma #IS879), 1 µM dexamethazone (Sigma #D4902), and 2 µM Rosiglitazone (Sigma #R2408), followed by 3% FBS and 1 µg/mL insulin for another 7 days, then maintained in 3% FBS for 8–11 days. After differentiation, cells were treated with diluted peptide and 100 nM insulin as positive control overnight in DMEM low glucose medium (Gibco #A1433001). Glucose uptake was measured by Glucose Uptake-Glo™ assay kit (Promega #J1343) according to the manufacturer’s recommendation. Briefly, cells were incubated with 0.5 mM 2DG in PBS for 10 min at 25 °C, before stop buffer and neutralization buffer were added sequentially. 2DG detection reagent were added to cell samples and incubated for 1 h before luminescence measurement on Envision (PerkinElmer). Data was analyzed with GraphPad Prism v7.

Glucose stimulated insulin secretion assay
The INSIE luciferase cells developed in the lab of Claes Wollheim, was transfected with the proinsulin-Luciferase reporter construct23. INSIE Luciferase stains21 cells were seeded in pre-coated 96-wells at 50,000 cells/well and cultured for 3 days in RPMI1640 medium with Glutamax (Gibco #61870), 1% Penicillin/Streptomycin (#15140-122), 10 µg/mL L-glutamine (#25030-081), 25 mM NaHCO3 (#25080-094), 0.2% BSA (SIGMA #B2064), 1% Penicillin/Streptomycin (#J1343) according to the manufacturer recommendation. In brief, cells were incubated at 37 °C for 30 min. After treatment the cells were incubated at 37 °C for 2 h. Cells were subsequently washed with KRBH buffer with diluted peptides in 16.7 mM HEPES (GIBICO #15630). cells were washed with KRHB buffer (1x Krebs-Ringer buffer) supplemented with 10 mM HEPES (#15630-080), 1% Glutamax (#35030-061), 24 mM NaHCO3 (#25080-094), 0.2% BSA (SIGMA #B2064), 1% Penicillin/Streptomycin (#15140-122) and adjusted to pH 7.4. 100 µL KRHB buffer with 0.5 mM glucose were added and incubated at 37 °C for 2 h. Cells were subsequently washed with KRHB buffer once before 100 µL KRHB with diluted peptides in 16.7 mM glucose were added at 37 °C for 30 min. After treatment the cells were centrifuged at 100 g for 2 min. 50 µL supernatant was transferred to a new 96-well plate. Nano-Glo luciferase assay (Promega #NI120) were performed according to manufacturer’s recommendation. In brief, Nano-Glo luciferase buffer was mixed with substrate at 50:1 ratio, before the mixture was added to cell supernatant at 1:1 ratio. The mixture was incubated at room temperature for 10 min before luminescence measurement on Envision (PerkinElmer). Data was analyzed with GraphPad Prism v7.

Hydrodynamic gene delivery
CHG plasmids and control plasmids were manufactured by Genscript (Nanjing, China) at transfection grade with endotoxin <25 EU/mg, and all plasmids were diluted in saline to a concentration of 20 µg/mL, 2.5 mL of pre-warmed dose solutions with 50 µg construct or 10 µg as positive control were carefully injected into conscious mouse through tail vein within 5 s using a 25 G syringe. Plasma was collected on termination day.

Plasma insulin measurement
Plasma insulin measurement was performed on Envision (PerkinElmer). Data was analyzed with GraphPad Prism v7. Plasma β–3-hydroxybutyrate (3-HB) measurement
3-HB were measured using a Cyclic Enzymatic Method and an auto bioanalytical analyzer Cobas C501 with reagents (Cat no. 4177350/4173630/41273791) purchased from Wako Chemicals GmbH, D-41468 Neuss.

Plasma triglycerides measurement
Triglycerides were measured by auto bioanalytical analyzer Cobas C501 with reagents (Cat no. 20767107322) purchased from Roche Diagnostics GmbH, D-68298 Mannheim.

Reporting summary
Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
Source data are provided with this paper as Supplementary Data files 1–6 and in the Source Data file. The raw mass spectrometry data and processed search files have been deposited at the ProteomeXchange Consortium via the PRIDE partner repository with the data set identifier PXD022225. Public peptide databases such as Sweek (http://www.swpep.org/), Uniprot (https://www.uniprot.org/) and NeuroPep (http://jsyslab.info/NeuroPep/) were used for Supplementary Data 3.

Code availability
The PPV code is available on the GitHub repository at https://github.com/jancr/ppv, released under the MIT licence (https://zenodo.org/record/7140868#.YOEvSNhbw2w). All plots can be regenerated from the Jupyter notebook: https://github.com/jancr/ppv/blob/master/notebooks/manuscript_figures.md.

References
1. Muttenzhaler, M., King, G. F., Adams, D. J. & Alewood, P. F. Trends in peptide drug discovery. Nat. Rev. Drug Discov. 20, 310–325 (2021).
2. Macdonald, P. E. et al. The multiple actions of GLP-1 on the process of glucose-stimulated insulin secretion. Diabetes 51, 434–442 (2002).
3. Young, A. A. Amylin regulation of fuel metabolism. J. Cell Biochem. 55, 12–18 (1994).
4. Lundberg, J. M. et al. Localization of peptide YY (PYY) in gastrointestinal endocrine cells and effects on intestinal blood flow and motility. Proc. Natl Acad. Sci. USA. 79, 4471–4475 (1982).
5. Sandoval, D. A. & D’Alessio, D. A. Physiology of proglucagon peptides: role of glucagon and glp-1 in health and disease. Physiol. Rev. 95, 513–548 (2015).
6. Tagliabracci, V. S. et al. Dynamic regulation of FGF23 by Fam20C. Nat. Commun. 520, 645 nm after excitation by a 680 nm laser. The total measurement time per well is 210 ms including a 70 ms excitation time. The lower limit of quantification is 20 pM.

Plasma β–3-hydroxybutyrate (3-HB) measurement
3-HB were measured using a Cyclic Enzymatic Method and an auto bioanalytical analyzer Cobas C501 with reagents (Cat no. 4177350/4173630/41273791) purchased from Wako Chemicals GmbH, D-41468 Neuss.

Plasma triglycerides measurement
Triglycerides were measured by auto bioanalytical analyzer Cobas C501 with reagents (Cat no. 20767107322) purchased from Roche Diagnostics GmbH, D-68298 Mannheim.

Reporting summary
Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
Source data are provided with this paper as Supplementary Data files 1–6 and in the Source Data file. The raw mass spectrometry data and processed search files have been deposited at the ProteomeXchange Consortium via the PRIDE partner repository with the data set identifier PXD022225. Public peptide databases such as Sweek (http://www.swpep.org/), Uniprot (https://www.uniprot.org/) and NeuroPep (http://jsyslab.info/NeuroPep/) were used for Supplementary Data 3.

Code availability
The PPV code is available on the GitHub repository at https://github.com/jancr/ppv, released under the MIT licence (https://zenodo.org/record/7140868#.YOEvSNhbw2w). All plots can be regenerated from the Jupyter notebook: https://github.com/jancr/ppv/blob/master/notebooks/manuscript_figures.md.

References
1. Muttenzhaler, M., King, G. F., Adams, D. J. & Alewood, P. F. Trends in peptide drug discovery. Nat. Rev. Drug Discov. 20, 310–325 (2021).
2. Macdonald, P. E. et al. The multiple actions of GLP-1 on the process of glucose-stimulated insulin secretion. Diabetes 51, 434–442 (2002).
3. Young, A. A. Amylin regulation of fuel metabolism. J. Cell Biochem. 55, 12–18 (1994).
4. Lundberg, J. M. et al. Localization of peptide YY (PYY) in gastrointestinal endocrine cells and effects on intestinal blood flow and motility. Proc. Natl Acad. Sci. USA. 79, 4471–4475 (1982).
5. Sandoval, D. A. & D’Alessio, D. A. Physiology of proglucagon peptides: role of glucagon and glp-1 in health and disease. Physiol. Rev. 95, 513–548 (2015).
6. Tagliabracci, V. S. et al. Dynamic regulation of FGF23 by Fam20C. Nat. Commun. 520, 645 nm after excitation by a 680 nm laser. The total measurement time per well is 210 ms including a 70 ms excitation time. The lower limit of quantification is 20 pM.
31. Olsen, J. V. et al. Higher-energy C-trap dissociation for peptide substrates of prolyl oligopeptidase: classical versus non-classical bioactive peptides. CNS Neurol. Disord. Drug Targets 10, 319–326 (2011).

32. Hansson, K. T. et al. Expanding the cerebrospinal fluid endopeptidome. Proteomics 17, 1600384 (2017).

33. Färh, M. et al. SwePep, a database designed for endogenous peptides and mass spectrometry. Mol. Cell. Proteom. 5, 998–1005 (2006).

34. Wang, Y. et al. NeuroPep: a comprehensive resource of neuro-peptides. Database 2015, 1–9 (2015). (Oxford).

35. Troger, J. et al. Granin-derived peptides. Prog. Neurobiol. 154, 37–61 (2017).

36. Marksteiner, J. et al. PE-11, a peptide derived from chromogranin B, in the human brain. Neuroscience 91, 1155–1170 (1999).

37. Merkler, D. J. C-terminal amidated peptides: Production by the in vitro enzymatic amidation of glycine-extended peptides and the importance of the amide to bioactivity. Enzym. Microb. Technol. 116, 450–456 (1994).

38. Bernay, B. et al. Discovering new bioactive neuropeptides in the striatum secretome using in vivo microdialysis and versatile proteomics. Mol. Cell. Proteom. 8, 946–958 (2009).

39. Toshinai, K. et al. Neuroendocrine regulatory peptide-1 and –2 (NERP) inhibit the excitability of magnocellular neurosecretory cells in the hypothalamus. Brain Res. 1563, 52–60 (2014).

40. Grannings, A. G. B., Kacprowski, T. & Ściedla, C. Multiplet: a hierarchical deep learning approach for multi-label classification of peptide bioactivities. Biol. Methods Protoc. 6, 1–16 (2021).

41. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. Nature. Genet. 25, 25–29 (2000).

42. Pletscher-Frankild, S., Palleja, A., Tsafou, K., Binder, J. X. & Jensen, L. J. DISEASES: text mining and data integration of disease-gene associations. Methods 74, 83–89 (2015).

43. Sheils, T. K. et al. TCRD and Pharos 2021: mining the human proteome for disease biology. Nucleic Acids Res. 49, D1334–D1346 (2021).

44. Karlsson, E., Stridsberg, M. & Sandler, S. Chromogranin-B regulation of IAPP and insulin secretion. Regul. Pept. 87, 33–39 (2000).

45. Yanibeygi, H., Maleki, M., Sathyapalan, T. & Sahebkar, A. The effect of C-peptide on diabetic nephropathy: A review of molecular mechanisms. Life Sci. 237, 116950 (2019).

46. Patel, V. K. et al. The role of C-peptide in the regulation of insulin sensitivity. Mol. Cell. Proteom. 4, 339–350 (2017).

47. Khatri, N. et al. Plasma gelsolin levels decrease in diabetic state and range from non-classical neuropeptides to microproteins. J. Proteome Res. 20, 11408–11419 (2021).

48. Madsen, T. D. et al. An atlas of O-linked glycosylation on peptide peptides. J. Proteome Res. 20, 11403–11415 (2021).

49. Kelstrup, C. D. et al. Rapid and deep proteomes by faster sequencing of secreted products from pancreatic islet culture results in a higher yield of full-length peptide hormones than found using cell lysis procedures. J. Proteome Res. 12, 3610–3619 (2013).

50. Secher, A. et al. Analytic framework for peptidomics applied to large-scale neuropeptide identification. Nat. Commun. 7, 1143 (2016).

51. Parker, B. L. et al. Multiplexed temporal quantification of the exercise-regulated plasma peptidome. Mol. Cell. Proteom. 16, 2055–2068 (2017).

52. Larraufie, P. et al. Important role of the GLP-1 axis for glucose homeostasis after bariatric surgery. Cell Rep. 26, 1399–1408 (2019).

53. Mooney, C., Haslam, N. J., Holton, T. A., Pollastri, G. & Shields, D. C. PeptideLocator: prediction of bioactive peptides in protein sequences. Bioinformatics 29, 1120–1126 (2013).

54. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. Nature. Genet. 25, 25–29 (2000).

55. Pérez-Riverol, Y. et al. The ProteomeXchange consortium in 2020: enabling ‘big data’ approaches in proteomics. Nucleic Acids Res. 48, D1154–D1152 (2020).

56. Perré, D. et al. The PRIDE database and related tools and services in 2019: improving support for quantification data. Nucleic Acids Res. 47, D442–D450 (2019).

57. Salvetat, J., Wiecki, T. V. & Fonnesbeck, C. Probabilistic programming in Python using PyMC3. PeerJ Comput. Sci. 2, e55 (2016).

58. Kullback, S. & Leibler, R. A. On information and sufficiency. Ann. Math. Stat. 22, 79–86 (1951).

59. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).
The authors declare the following competing interests: The authors are or have been working for Novo Nordisk A/S, a pharmaceutical company with commercial interest in bioactive peptides for the treatment of diabetes and obesity. All authors except J.C.R, F.G.T, U.L, and J.V.O. hold minor share portions as part of their employment in Novo Nordisk. J.V.O. consults for and has research funding from Novo Nordisk.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-34031-z.

Correspondence and requests for materials should be addressed to Christian T. Madsen.

Peer review information Nature Communications thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022