**A proteome-wide screen defines binding determinants of the core autophagy protein LC3B**

Jennifer Kosmatka1, Cong Liu2, Daniel Lim1, Joseph H. Davis1,3 \*, Amy E. Keating1,4 \*

* Need to add Jackson

1 Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139, United States.

2 Broad Institute of MIT and Harvard, Cambridge, Massachusetts, 02139, United States.

3 Computational and Systems Biology Graduate Program, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139, United States.

4 Department of Biological Engineering and Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139, United States.

\* Correspondence to: [jhdavis@mit.edu](mailto:jhdavis@mit.edu) or [keating@mit.edu](mailto:keating@mit.edu)

Classifications: Major – Biological Sciences; Minor – Biochemistry

Keywords: autophagy, short linear motifs, LIR, LC3B, bacterial display

**ABSTRACT [223/250 word max]**

Human MAP1LC3B (LC3B) binds proteins involved in autophagy and other cellular processes using a degenerate four-residue short linear motif known as the LC3-interacting region (LIR). Biochemical and structural studies have identified LIRs in many LC3B interaction partners, but the sequence features that contribute to binding have not been systematically explored. To discover peptides that interact with LC3B and deeply profile the key binding determinants, we screened a library of ~500,000 36-amino acid peptides derived from the human proteome using bacterial cell-surface display. Analysis of the screening data coupled with structural studies and site-directed mutagenesis revealed exceptions to the reported LIR motif and a strong preference for negatively charged residues adjacent to the LIR. Using peptide binders from the screen as a guide, we designed synthetic LIR-containing peptides that bind LC3B with affinities comparable to the tightest measured natural binder. We constructed a PSSM composed of the highly enriched peptides to accurately predict whether peptide sequences are likely to bind LC3B, offering more information content than previous methods. Finally, we determined that mutations in LC3B commonly thought to abrogate binding of LIR-containing peptides instead alter LC3B binding specificity, leading to enhanced binding of some LIR sequences. Taken together, our results refine the LIR motif definition, expand the network of candidate LC3B interaction partners, and enable the prediction of LC3B-interacting LIRs.

**SIGNIFICANCE STATEMENT [120 words – undergraduate scientist level]**

Short linear interaction motifs in disordered regions of proteins play crucial roles in mediating interactions with folded protein domains throughout biology. Core autophagy protein LC3B interacts with many distinct binding partners throughout the autophagic process using a single docking site that binds to proteins bearing a cognate short linear motif known as the LC3-interacting region, or LIR. To understand the sequence features that contribute to these interactions, we used a high-throughput screen to identify peptides from the human peptidome that bind to LC3B. Our results more precisely define the LIR motif, identify new candidate interaction partners, and support a model that can predict whether human LIR-containing peptides bind to LC3B.

**INTRODUCTION**

Intrinsically disordered regions (IDRs) in proteins participate in key cellular processes via interactions with globular protein domains (Dyson and Wright 2005). Short linear motifs (SLiMs) mediate these interactions using contiguous stretches of 3 – 10 amino acids that undergo disorder-to-order transitions upon binding, typically resulting in interactions with equilibrium dissociation constants (KD values) in the micromolar range (Tompa et al. 2014; Wright and Dyson 2015; Diella 2008; Davey et al. 2012; Van Roey et al. 2014). For most SLiMs, the sequence features necessary and sufficient for binding are incompletely understood. Cataloged regular expressions, such as those in the Eukaryotic Linear Motif (ELM) and Motif Map of the Proteome (MoMaP; https://slim.icr.ac.uk/momap/) databases, (Kumar et al. 2022) capture the features that best typify known binders but also highlight the limitations of existing motif specifications, which are satisfied by many sequences that fail to bind (Bugge et al. 2020). Whereas the number of experimentally defined SLiMs has expanded in recent years through the targeted assessment of individual binding partners, more systematic techniques are required to refine these recognition patterns and to support the development of prediction tools (Kumar et al. 2024; Davey et al. 2023).

SLiMs mediate interactions involved in critical cellular processes, including the conserved proteostasis pathway macroautophagy (hereafter autophagy). Autophagy involves the *de novo* formation of a double-membrane organelle known as the autophagosome, which transports cellular contents to the lysosome for degradation (Mei et al. 2014; Popelka 2020). Microtubule-associated protein 1A/1B light chain 3B (LC3B) and its homologs are critical to every stage of the autophagic process and participate via interaction with numerous partners that contain a SLiM termed the LC3-interaction region (LIR) (Rogov et al. 2023)(Birgisdottir et al. 2013). During autophagy, hundreds of LC3B proteins are covalently linked to the growing autophagosome via reversible conjugation to phosphatidylethanolamine embedded in the autophagosomal membrane (Schmitt et al. 2022). LC3B-LIR complexes, enhanced by multivalency and avidity at the membrane surface (Sawa-Makarska et al. 2014; Wurzer et al. 2015; Lee and Davis 2024), tether cellular components to the autophagosome either directly as cargo or indirectly using selective autophagy receptors to ensure their efficient degradation (Rogov et al. 2014). LC3B-LIR interactions are also critical for autophagosome biogenesis, transport, and lysosomal fusion (Johansen and Lamark 2020). Mutations that alter LC3B-LIR interactions have been linked to the development of many diseases, including neurodegenerative disorders, aging, and cancer (Park et al. 2020; Brennan et al. 2022; Ramesh Babu et al. 2008; Fas et al. 2021; Chen et al. 2024). LC3B also has important roles in non-autophagic processes, most notably LC3-associated phagocytosis (LAP) (Florey et al. 2011), LC3-associated endocytosis (LANDO) (Heckmann et al. 2019), and LC3-associated micropinocytosis (LAM) (Galluzzi and Green 2019).

The hAtg8 proteins (consisting of the LC3 and GABARAP subfamilies) share a common architecture consisting of a ubiquitin-like beta-grasp fold with two additional N-terminal alpha-helices (Rogov et al. 2023; Noda et al. 2008). There are many examples of the 14 kDa LC3B and its paralogs binding to the information-poor consensus LIR [FWY]0-X1-X2-[LVI]3 (Chatzichristofi et al. 2023), where X can be any amino acid, via two conserved, hydrophobic pockets that form the LIR-docking site (LDS). In its canonical binding mode, the aromatic residue [FWY]0 engages the first pocket (HP1), and the aliphatic residue [LVI]3 engages the second pocket (HP2) (Rogov et al. 2023). This degenerate LIR motif occurs ~170,000 times in the human proteome, with ~19,000 of such occurrences falling in regions of predicted disorder. It is unlikely that all 19,000 of these sequences represent functional interaction sites and, indeed, there are annotated examples of proteins containing canonical LIRs that fail to bind LC3B with measurable affinity (Chatzichristofi et al. 2023). Moreover, in recent years, the discovery of binders that do not match the canonical LIR motif has expanded the number of known hAtg8 binding modes, and thus the number of potential binders. For example, Li et al. reported that LC3B binds to ankyrin-G using the LIR sequence 1989WIEF1992, wherein an aromatic Phe residue engages HP2 (Li et al. 2018), and Knævelsrud and colleagues found that sorting nexin 18 (SNX18) binds hAtg8 paralogs using an expanded five-residue motif (154WDDEW158) (Knævelsrud et al. 2013). The proteome-wide abundance of such non-canonical binders, i.e., binders that do not match the motif [FWY]0-X1-X2-[LVI]3, is not known.

Given that not all instances of the LIR motif bind to LC3B, and that not all LC3B interactors contain a LIR motif, we sought to understand additional determinants beyond the core LIR that contribute to binding. To discover these, we used a bacterial display assay to screen ~500,000 peptides derived from the human proteome for those capable of binding to LC3B. We determined the binding affinity of top hits from the screen and elucidated binding mechanisms that support SLiM-LC3B interactions for known and novel binders through structural and biochemical analysis of identified peptides and mutated variants. Our results expand the number and types of residues that can engage HP2 on LC3B and reveal the influence of N-terminal acidic residues on LIR·LC3B binding affinity. Our results allowed us to design a synthetic LIR peptide with affinity comparable to that of the tightest known LC3B binder. Furthermore, using our screening data, we created a PSSM to identify peptides likely to bind LC3B.

**RESULTS**

**A high-throughput bacterial display screen identifies thousands of LC3B-binding peptides.**

To rapidly identify hundreds of peptides capable of binding LC3B, we used bacterial surface display in combination with fluorescence-activated cell sorting (FACS). Following Hwang *et al.*, we expressed ~500,000 36-mer peptides spanning the human proteome on the cell surface of *Escherichia coli* via fusion to circularly permuted OmpX (Daugherty 2007; Larman et al. 2011; Hwang et al. 2022). To compensate for the low micromolar affinity of typical SLiM interactions, and to better mimic the avid interactions enabled by the high local concentration of LC3B conjugated to autophagosomal membranes in cells(Zaffagnini and Martens 2016), we bound N-terminally biotinylated LC3B to tetravalent streptavidin-phycoerythrin (SA-PE). Using FACS, we quantified peptide expression levels using an allophycocyanin (APC)-conjugated antibody against an encoded N-terminal FLAG tag. Cells displaying peptides that bound to LC3B were identified and isolated by their high SA-PE / APC ratio (**Figure 1A**). We validated our screening approach using well-characterized LC3B-binding peptides of various affinities (**Supplemental Figure 1**). Following five rounds of positive selection and one round of counterselection against non-specific binding to SA-PE (**Figure 1B**), LC3B-binding sequences were enriched ~80-fold, from ~0.7% of the naïve library, to ~57% of all expressing cells (**Supplemental Figure 2**).

In all, 12,158 peptides from 5,578 unique proteins were identified in the final round of library sorting, with individual peptides displaying a wide range of behaviors across the sorts, as assessed by their calculated enrichment ratio (ER) in each sort (**Figure 1C**, see Methods) (Rubin et al. 2017). To select the highest enriching binders for further analysis and validation, we identified those with an average ER more than 1.70 standard deviations above the mean (z 1.70) over the last three rounds of sorting **(Figure 1D-E**, n=427 peptides), which is a threshold satisfied by known high-affinity binders.

Our input library contained seven peptides with LIR motifs previously shown to bind LC3B with sub-5 µM affinity (Chatzichristofi et al. 2023). Of these, four were captured throughout our six rounds of sorting, all exhibiting ER z-scores ≥ 1.70 (**Figure 2A, Supplemental Figure 3**). TBCD5 and SNX18, though unmeasured for LC3B affinity, show high enrichment in the sort. Of three additional known binders in the library with affinities between 5 and 50 µM, one was carried through six rounds of sorting but exhibited an ER z-score < 1.70. Those that were not captured were either selected against in the negative sort or had weak affinities for LC3B. One example is ULK1 (KD = 48 µM) (Wirth et al. 2019), which did not pass the screening criteria, likely due to the limit of detection in bacterial display (**Supplemental Figures 1** and **3**). Enrichment behavior for peptides with measured affinity < 50 µM for LC3B are shown in Supplemental Figure 3.

The screen also identified binding peptides from sixteen proteins previously reported to co-immunoprecipitate with LC3B, the majority of which lacked confirmed interaction sites prior to this study (Stark 2006) (**Figure 2A**). We tested four such peptides derived from MAP1B823-858, ATG4A363-398, SCYL1640-675, and HEAT3377-412 for binding to soluble, monomeric LC3B by bio-layer interferometry (BLI) and observed robust binding with dissociation constants ranging from ~0.5 to 10 µM (**Figure 2A, B, C, Table 1**). Notably, all four peptides contain canonical LIR motifs and acidic residues N-terminal of the LIR, suggesting the importance of these elements in supporting LC3B binding.

**Strongly enriched peptides include novel, candidate LC3B interactors.**

To identify peptide hits from our screen (z ≥ 1.70) that are most likely to represent biologically relevant interaction partners, we prioritized those corresponding to proteins that share GO annotations (Huntley et al. 2015) with LC3B (**Figure 2A**, **Supplemental Figure 4**). We tested LC3B for binding to HERC13075-3110 and LRRK2858-893, noting that these proteins have been reported to impact autophagy (Montes-Fernández et al. 2020; Pérez-Villegas et al. 2020; Boecker et al. 2021; Park et al. 2016; Roosen and Cookson 2016; Madureira et al. 2020) but that direct interactions with LC3B have not been reported. Each peptide bound with a dissociation constant in the low micromolar range (**Figure 2C**, **Table 1, Supplemental Table 1**). We additionally identified a subset of sequences amongst our peptide hits that were derived from proteins lacking any shared GO terms with LC3B or any sequences previously identified to bind to interact with LC3B. This set of sequences, which we refer to hereafter as “Unannotated”,nonetheless contained multiple sequences that bound LC3B with dissociation constants ≤ 50 µM affinity and thus represent candidates for new LC3B-interacting proteins (**Figure 2C, Table 1,** **Supplemental Table 1**).

**Enriched LIR-containing sequences reveal a preference for Trp in X0 preceded by acidic residues.**

To identify sequence features common to the highest enriching LIR-containing peptides, hits (z ≥ 1.70) were analyzed using pLogo (O’Shea et al. 2013), with the LIR-containing fraction of the input library as background (**Figure 3A**). Consistent with this logo capturing features related to binding affinity, a consensus peptide derived from the residues indicated as most over-enriched in binders, pCONSLIR bound to LC3B with a dissociation constant of ~60 nM (**Figure 3B**). This affinity is on par with the tightest known LC3B-binding peptide, Ank21578-1613 (Li et al. 2018a). pCONSLIR alsobinds~30-fold tighter than a well-studied peptide from FYCO11277-1312 (Olsvik et al. 2015; Cheng et al. 2016) and is ~25-fold higher in affinity than a chimeric peptide designed to target alpha-synuclein for autophagic degradation (Tong et al. 2023).

The LIR-containing pLogo exhibited significant enrichment of tryptophan in the first position (X0), over tyrosine and phenylalanine, and a preference for acidic resides in positions X-1, X-2, and X-3, as well as for glutamate in the X1 of the core LIR motif and at the C-terminal X7 position (**Figure 3A**). We partitioned the sequence alignments based on the W/F/Y residue identity at the X0 position and generated three new peptides based on the consensus sequences (pCONSW, pCONSY, pCONSF; **Supplementary Figure 6)**. Peptide pCONSW bound to LC3B with a dissociation constant of ~90 nM, nearly as tight as the consensus pCONSLIR (**Figure 3B, C)**. We tested the contributions of the highly enriched N- and C-terminal acidic residues through analysis of this pCONSw peptide. We found that whereas the C-terminal truncation had a modest effect on affinity, the N-terminal truncation decreased affinity 10-fold (**Figure 3C**,row 1 vs. 3 and 7). Most of this reduced affinity could be restored by re-introducing either a single N-terminal Asp or a 6-residue EDDDDA sequence that lacked this Asp (**Figure 3C**, row 1 vs. 5 and 6). These data support an important role for flanking N-terminal acidic residues in achieving high-affinity binding to LC3B, without a strict requirement for a specific residue at a specific site. Consistent with the key contributions being proximal to the LIR, removing 7 residues from both the N and C-terminus of the 36-residue peptide identified in the screen had no impact on affinity (**Figure 3C**, row 1 vs 2 and 4).

**Acidic residues N-terminal of the core LIR enhance interactions with LC3B through structural contacts.**

Many of our top-scoring peptides exhibited features common to both pCONSLIR and pCONSw, namely the presence of Trp in X0, Glu in X1, and multiple N-terminal acidic residues. Indeed, two of the top-scoring peptides (BLM552-570\* DIDNFDIDDFDDDDDWEDI*CII\**, z-score = 4.1, rank = 1; and BLM552-587 DIDNFDIDDFDDDDDWEDIMHNLAASKSSTAAYQPI, z-score = 3.6, rank = 12) contained these PSSM-enriched features, and each bound tightly to LC3B (KD ~ 0.7 µM). Given that many high-affinity LC3B binders, including FYCO11276-1288, ANK21588-1613, and ANK31985-2010, use an acidic residue in position X7 of the C-terminal extension to make an affinity-enhancing contact to LC3B residue R70 (Olsvik et al. 2015; Li et al. 2018b; Cheng et al. 2016), which these sequences lack, we sought to understand how these BLM-derived peptides bound so tightly to LC3B. We determined the structure of BLM552-571 bound to LC3B to a resolution of 2.2 Å using X-ray crystallography (**Figure 3D**, **Supplemental Table 2**). The interactions between the peptide and LC3B in this structure are consistent with those observed for other canonical LIR motifs, with the aromatic residue W567 (X0) and hydrophobic residue I570 (X3) deeply engaged in HP1 and HP2 (**Figure 3F, G**). Additionally, an intermolecular beta sheet forms between the main chain of LC3B residues F52 and L53 and BLM residues W567 (X0), E568 (X1), and I570 (X3). Interestingly, the sidechain of BLM E568 (X1) contacts LC3B R70, the same residue contacted by known C-terminal extensions (PDB: 5D94 Olsvik et al. 2015; PDB: 5YIQ, 5YIS Li et al. 2018a; PDB: 5CX3 Cheng et al. 2016) (**Supplemental Figure 7**), and the N-terminal acidic residues D565-D567 (X-3 – X-1) are in close proximity to positively charged side-chain atoms of L3CB R11 and K51 (**Figure 3E**), further emphasizing the contributions of acidic residues N-terminal to the LIR.

**Many highly enriched peptides lack a canonical LIR motif.**

Nearly half of the 427 highest-enriching peptides lack a canonical LIR motif. Whereas seven of these peptides contain a biotin-mimicking HPQ motif that can bind to streptavidin (Devlin et al. 1990; Weber et al. 1992) (**Supplemental Figure 8**), over half of the remaining peptides contained a sequence matching [FWYLVI]0-X1-X2-[LVIFWY]3 (**Supplemental Figure 9**), which we hypothesized might support LC3B binding. In this paper, we will refer to the previously reported motif [FWY]0-X1-X2-[LVI]3 as the “canonical LIR” and the more permissive motif [FWYLVI]0-X1-X2-[LVIFWY]3 as “LIR+” as this redefinition reflects an expanded residue promiscuity. To determine whether this superset of the canonical LIR can support binding, we selected three highly enriching peptides bearing LIR+ sequences: DYH12422-445\* with two LIR+ sequences; PAR118-37 with a single LIR+ sequence, and CTSL2233-262 with both a canonical LIR and a LIR+ motif.

In each instance, alanine substitutions of the LIR+ sequence reduced affinity for LC3B, consistent with a critical role for LIR+ sequences in supporting LC3B association. Notably, in CTSL2233-262, mutation of the LIR+ sequence 246WEVF249 to AAAA produced a much stronger effect than mutation of the canonical 252YYFI255 LIR motif to AAAA (**Figure 4A**). Although sequence variations of the canonical LIR motif exist (Li et al. 2018b; Ibrahim et al. 2023), our results illustrate that such motifs may be a more common mode of engagement than previously appreciated.

**A model trained using a high-throughput interaction screen predicts LC3B-binding peptides.**

Having established both the importance of the core and N-terminal flanking sequence features with our previous analysis and the important role of LIR-adjacent motifs binding LC3B, we next sought to develop a method to assess whether the combination of LIR- and LIR+- containing sequences with unknown affinity for LC3B were likely to bind this target. To achieve this, we trained a series of classifier models (see Methods) using our highly enriched (z-score ≥ 1.70) LIR- and LIR+-motif peptides as positive examples. Peptides that consistently depleted across sorts 1 and 3 before dropping out were considered nonbinders. We found each model was similarly performant, with a ROC AUC of ~0.75 when assessed using a held-out test-set derived from the LIRCentral database (Chatzichristofi et al. 2023). Notably, these models outperformed (p<X) the current state-of-the-art prediction server, iLIR, which exhibited a ROC AUC of **Z** on this test set. Interestingly, the performance was robust to an extensive series of model perturbations, including the changes in overall model architecture and hyperparameters, indicating that the incorporation of our unbiased screening data, and not new modeling frameworks, led to the observed improvement in performance. We validated the robustness of the model by predicting the binding class of all sorting peptide hits confirmed through BLI (**Supplemental Figure X)**. We found the model distinguished which LIR or LIR-like motif drove binding to LC3B in peptides containing multiple LIR motifs (**Supplemental Figure X**).

**LC3B pocket mutations F52A and L53A modulate but do not ablate peptide binding.**

LC3B with mutations F52A and L53A, here called LC3B LDS\*, has been used to assess binding at the LDS. These mutations, in HP2 and HP1, respectively, have been assumed to disrupt peptide binding at this interface (**Figure 4B**) (Marshall et al. 2019; Kraft et al. 2014; Behrends et al. 2010; Skytte Rasmussen et al. 2017; Qiu et al. 2017). Using LC3B LDS\*, we performed three additional sorts with the library from sort 5 as the input (**Supplemental Figure 9**, see Methods). A well-studied control peptide from FYCO1, FYCO11277-1312, binds seven-fold weaker to LC3B LDS\* than to LC3B and was depleted throughout the three rounds of LDS\* sorting (**Figure 4C-D**), consistent with its canonical LDS-dependent binding mechanism. Similar behavior was observed for the other known LC3B-binding peptides, including KBTB7639-674, FUND11-36, NEDD4681-716, and ATG4A363-398 (**Figure 4C-D**). Notably, however, a subset of peptides was further enriched when sorted for binding to LC3B LDS\* (**Figure 4C-D**).

We selected six peptides enriched for binding to LC3B LDS\* and quantified their binding to wild-type LC3B or LC3B LDS\*. These peptides bound LC3B LDS\* with low micromolar affinity and with variable affinity for the wild-type LC3B. Specifically, we determined that peptides can: 1) exhibit reduced affinity towards LC3B LDS\* compared to wildtype (FYCO11277-1312 and ATG4A363-398); 2) exhibit no preference (PPM1H436-471); or 3) bind preferentially to LC3B LDS\* (OSBL71-36, TRIM5288-120, PEP2) (**Figure 4D, Supplementary Table 1)**. A structural model of LC3B LDS\* shows that the binding interface is altered, but the hydrophobic pockets persist (**Figure 4B**). Thus, it is perhaps unsurprising that these commonly used mutations differentially modulate the binding of LIR- and LIR+ peptides by creating a new hydrophobic interface that is preferred or disfavored, depending on the specific SLiM (**Figure 4B)**

**DISCUSSION**

LC3B and other hAtg8 paralogs can interact with binding partners via the core LIR motif [FWY]0-X1-X2-[LVI]3,an information-poor SLiM that forms critical interactions in all major stages of bulk and selective autophagy and in non-autophagic processes such as LAP and LANDO (Rogov et al. 2014; Florey et al. 2011; Heckmann et al. 2019; Galluzzi and Green 2019). In this study, we used high-throughput bacterial-surface-display to broadly assess the LC3B-binding potential of peptides derived from the human proteome, with the goal of better detailing the interaction determinants of LC3B. We focused our analysis on the most strongly enriched peptides, recognizing that we would miss many authentic but weakly interacting partners. Interestingly, within this highest enriching set, ~40% of the identified peptides lack a traditional LIR but contain at least one non-canonical “LIR+” motif of the form [WFYILV]0-X1-X2-[WFYILV]3 where the canonical LIR is excluded. Although such motifs have been previously reported (Li et al. 2018), their prevalence was underappreciated: Only 0.98% of cataloged hAtg8 binding peptides contain the motif [FWY]xx[FWY] (Chatzichristofi et al. 2023). We additionally uncovered 4 peptides bearing a variation on a LIR motif in which the core hydrophobic residues were separated by either 3 or 4 acidic amino acids. Such altered registers have been reported (Knævelsrud et al. 2013) and identified in our screen, suggesting the possibility of them binding to LC3B. Our observations extend the allowable SLiM sequence space supporting binding to LC3B family members and suggest that additional determinants, such as local structure likely define the extent of peptide engagement by LC3B.

Using an unbiased Motif Discovery and Enrichment Analysis (XSTREME) (Grant and Bailey 2021) of our highest enriching peptide set, we identified the high-affinity motif [DE]-[ED]-[DE]-[WFY]-[E]-X-[LIVFM], which highlights the generally beneficial impact of acidic residues in N-terminal positions X-3 through X-1 and aligns with previous studies of individual peptides derived from FYCO1 and SQSTM1, for which LC3B binding is potentiated by N-terminal acidic residues (Pankiv et al. 2007). Further, through biochemical binding assays that tested the effects of mutations targeting this region, we found that N-terminal acidic residues generally improved binding affinity, without a strict position-specific dependence. We note that enrichment of acidic residues in the highest affinity binders is consistent with a reported affinity-enhancing role for phosphorylated serine and threonine residues in the N-terminal flanking sequence of the core LIR motif (Wirth et al. 2021; Richter et al. 2016; Kliche et al. 2023). Indeed, given this observed beneficial impact of N-terminal acidic residues, it is plausible that phosphorylation of target proteins could tune LIR-LC3B binding affinity, effectively linking kinase-dependent environmental signaling cascades to the regulation of autophagic cargo selection (**Figure 4B**) (Kliche and Ivarsson 2022; Birgisdottir et al. 2013; Rogov et al. 2023).

Recently, the Viestra group reported an alternate binding interface on Atg8 homologs (hAtg8) that they termed the UDS, or Ubiquitin-interaction-motif Docking Site. This interface, which occurs on the face opposite that of the LIR docking site, is reported to interact with an alternative SLiM, defined as the UIM, with the sequence Ψ-ζ-X-A-Ψ-X-X-S, with Ψ, ζ and X representing small hydrophobic residues, hydrophilic residues, and any amino acid, respectively (Marshall, et al. 2019). We did not observe enrichment for peptides bearing this motif in our screen, indicating that if such an interface exists, it was either unavailable for binding in our screening platform or did not support sufficiently high-affinity interactions to be captured by our screen. Notably, we did not observe enrichment for UIM-bearing peptides even when screening in the context of LC3B constructs bearing mutations at the LIR docking site (LDS\*). Instead, screening against LC3B LDS\* revealed that peptides bearing canonical LIR sequences exhibited altered affinity for the mutant LC3B domain, consistent with these often-utilized mutations altering but not ablating binding, contrary to what has been assumed (Marshall et al. 2019; Kraft et al. 2014; Behrends et al. 2010; Skytte Rasmussen et al. 2017; Qiu et al. 2017).

LIR databases exist that catalog experimentally validated motifs (Chatzichristofi et al. 2023) and use an extended six-residue LIR PSSM to predict LIR binders to hAtg8s and other orthologs (Kalvari et al. 2014). Our work provides further improvement in predicting binders solely to LC3B and this dataset allowed us to design a synthetic peptide that binds as tightly as anything previously reported. While we provide all high confidence predicted binders for the human proteome (Supplemental File X), we postulate that this prediction model can extend to other proteomes, e.g. viruses, that may result in the prediction of additional proteins that interact with LC3B in autophagy.

Together, the LC3B and LC3B LDS\* bacterial display sorts presented here provide a valuable resource for the study of SLiMs and autophagy. These datasets offer large-scale, high-resolution insight into preferred flanking residues surrounding the canonical LIR motif, expand the repertoire of LC3B-binding sequences, and highlight important caveats in the use of the LC3B F52A L53A mutant to disrupt LDS-dependent interactions. While these data provide high-resolution insight into LC3B binding preferences, the bacterial display system may not fully recapitulate the complexity of LC3B interactions in mammalian cells because of required avidity, localization, and post-translational modifications common to SLiMs. In addition, our findings enhance the ability to predict LIR–LC3B interactions across the proteome and nominate new candidate LC3B interactors, laying the groundwork for future functional studies.

**ACKNOWLEDGEMENTS**

We thank A. Ghandanpour for help with crystallographic data collection and structure refinement. We thank J. Halpin for help with bioinformatic code. We thank L. Kinman, D. Cui, and B. Powell for helpful discussions. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Numbers R35GM149227 (AK) and R01GM144542 (JHD), and the Smith Family Odessey Award (JHD).

**AUTHOR CONTRIBUTIONS**

J.E.K., J.H.D., and A.E.K. designed research; J.E.K. performed screening, data analysis, biochemical experiments and contributed new reagents and analytic tools; C.L., J.E.K., and J.H. developed model; D.L. collected and refined crystallography data; J.E.K., C.L., J.H.D., and A.E.K. wrote the paper.

**COMPETING INTERESTS**

The authors declare no competing interests.

**DATA AVAILABILITY**

Raw sequencing reads will be deposited upon publication. Crystal structure deposited to the PDB with accession number XXX. Python processing scripts will be made available upon request. All processed data are included in the manuscript and/or SI Appendix.

**MATERIALS AND METHODS**

**LC3B and LC3B LDS\* Expression and Purification**

For bacterial display, monomeric LC3B and LC3B LDS\* were cloned into a pDW363 biotinylation vector encoding an N-terminal Biotin Acceptor Peptide (BAP) and 6xHis tag; the plasmid also encodes ligase BirA. To express peptides for biolayer interferometry (BLLI) experiments, the BAP tag was removed. LC3B and LC3B LDS\* were expressed in Rosetta2(DE3) (Novagen) cells in 2xYT with 100 ug/mL ampicillin and 0.05 mM D-(+)-biotin (not added for BLI preps) at 37 °C and shaking at 225 RPM. Cells were induced with 1 mM final IPTG and grown at 25 °C overnight before harvesting by centrifugation at 5000xg for 15 minutes. Cell pellets, if not immediately processed, were flash-frozen and stored at -80 °C. Cell pellets were resuspended in NiNTA binding buffer (20 mM Tris-HCl pH 8.0, 500 mM NaCl, 5 mM imidazole) supplemented with 0.2 mM phenylmethylsulfonyl fluoride (PMSF) protease inhibitor. Cells were dounced 20 times, followed by sonication on ice 4 times for 2:30 mins at 30% amplitude, 20 seconds on, 10 seconds off. The lysate was centrifuged at 15,000xg for 15 minutes, followed by ultracentrifugation of the supernatant at 50,000 RPM for 1 hour. The clarified supernatant was loaded onto a 5 mL Bio-Scale Mini Nuvia IMAC Ni-charged cartridge (BioRad Cat#780-0811) pre-equilibrated with NiNTA binding buffer and eluted using a gradient of NiNTA elution buffer (20 mM Tris-HCl pH 8.0, 500 mM NaCl, 300 mM imidazole) with a final concentration of 300 mM imidazole. Target fractions were identified via SDS-PAGE, pooled, concentrated, loaded on a Superdex 75 16/600 pre-equilibrated column, and run in gel filtration buffer (20 mM Tris-HCl pH 8.0, 150 mM NaCl, 0.5 mM TCEP, 10% glycerol). Purity was verified via SDS-PAGE, and relevant fractions were concentrated, flash-frozen, and stored at -80 °C.

**Bacterial cell surface display plasmids, T7-Pep, and FACS Sample Preparation**

Control constructs for bacterial display were cloned at the C-terminus of circularly permuted OmpX (eCPX) (Rice and Daugherty 2008). Each peptide was expressed on the cell surface following an N-terminal FLAG tag and with a C-terminal cMyc tag. The T7-Pep library (a gift from the Elledge lab, Harvard University and Brigham and Women’s Hospital) was displayed in the same construct, as described by Hwang et al. (Hwang et al. 2022). FACS samples were prepared as previously described with slight modifications (Hwang et al. 2022). Briefly, the eCPX control or library plasmids were grown in electrocompetent MC1061 cells in 5 mL LB cultures with 25 ug/mL chloramphenicol and 0.2% w/v glucose rotating at 37 °C overnight. The next morning, the cells were inoculated in 5 mL TB with 25 ug/mL chloramphenicol and rotated at 37 °C until the OD600 reached 0.5-0.6. Cells were induced to a final concentration of 0.04% w/v arabinose and induced for 1.5 hours rotating at 37 °C. Following induction, 1 x 107 cells/sample or 7 x 107 cells/library were pelleted, washed, and resuspended in PBS + 0.1% BSA to reach a final concentration of 4 x 108 cells/mL. Cells were incubated with 1:100 anti-FLAG APC diluted in PBS + 0.1% BSA (PerkinElmer) at a ratio of 30 uL labeled antibody: 107 cells. 1.5 mL Eppendorf tubes were wrapped in foil and rotated at 4 °C for 30 minutes before being washed with 500 uL PBS + 0.1% BSA. While incubating, biotinylated LC3B was pre-tetramerized to streptavidin-PE (SAV-PE, ThermoFisher Scientific) at a 4.2:1 molar ratio in PBS + 1% BSA and 4 mM DTT (for a final concentration of 2 mM DTT). The protein was incubated for 15 minutes, rotating at 4 °C and wrapped in tin foil. To prepare the final sample, 25 uL of 1 x 107 cells in PBS was added to 25 uL of a 2x concentration of pre-tetramerized LC3B and incubated for 1 hour wrapped in tin foil, rotating at 4 °C. Samples were then transferred to a 0.22 micron 96-well Multi-Screen HTS GV sterile filtration plate (Millipore) pre-equilibrated with 500 uL PBS + 0.5% BSA. The cell solution was vacuum filtered, washed with 200 uL PBS + 0.1% BSA and resuspended in 250 uL PBS + 0.1% BSA for subsequent FACS analysis or sorting.

**Bacterial Display Sorting**

The naïve T7-pep human peptidome library was prepared as described in the FACS sample preparation section. A total of 7 x 107 cells were incubated with a final concentration of 1.68 µM pre-tetramerized LC3B and sorted on the BD FACSAria 4. To establish the appropriate gate for sorting, two positive controls (FYCO1 and ATG4A) and a negative (empty) control were analyzed. The gate was designed to maximize the number of positive cells, while minimizing the number of negative cells (0.1-0.4%) that were present in the gate. This gate was held constant through the enrichment process, following compensation. For the initial sort, enough cells were collected to oversample the naïve library by 100-fold. Following sorting, cells collected in SOC were rescued in LB containing 25 ug/mL chloramphenicol and 0.2% w/v glucose, rotated at 37 °C overnight, and harvested before the OD600 surpassed 1.0. The collected cells were then miniprepped to isolate the plasmid DNA. For the next round of sorting, 50 ng of plasmid DNA was desalted and transformed into electrocompetent MC1061s to eliminate potential growth biases within the bacteria. For the negative sort, biotin was pre-tetramerized to SAV-PE and a gate was selected that maximized the FACS profiles of the positive controls. The positive enrichment process was then repeated for sorts 3 through 6, each time collecting enough cells to oversample the expected binding pool by 100-fold.

The LC3B LDS\* sorting samples were prepared similarly, instead using 1.68 µM LC3B LDS\* pre-tetramerized to SAV-PE. For direct comparison, the binding pool from sort 5 was sorted against pre-tetramerized LC3B (Sort 6) and pre-tetramerized LC3B LDS\* (LDS\* 1 on the same day and both the binding and nonbinding populations were collected for sequencing. Two subsequent LC3B LDS\* enrichment sorts were performed to further enrich for peptides potentially binding an alternate site. For the LDS\* sorts, as well as canonical sort 5 and sort 6, both the binding and nonbinding populations were collected and miniprepped to isolate the plasmid DNA.

**NGS Sample Preparation**

As described above, cells from each round of sorting were grown overnight in LB containing 25 ug/mL chloramphenicol and 0.2% w/v glucose, rotating at 37 °C, then miniprepped (NEB Monarch). Samples collected on different days, or from sorts using different protein targets (LC3B or LC3B LDS\*) were assigned a unique index. We PCR amplified the variable region of the library with a forward primer containing the 5’ Illumina adaptor sequence (ol\_JD806) and a corresponding reverse primer that contained one of 14 6-nucleotide (nt) index sequences for subsequent multiplexing and the 3’ Illumina adaptor sequence (ol\_JD807-813, NGSi8-14); the amplicon preparation scheme and all primers are listed in Supplementary File 1. Eleven cycles of amplification were performed using Phusion polymerase (NEB) with an annealing temperature of 68 °C and 1.25 µM of the relevant primers. Two cycles of re-conditioning PCR with 1 µM primers and an annealing temperature of 68 °C were performed to eliminate potential heteroduplex formation. The resulting reactions were purified using a 0.85X/0.5X double-sided size selection with AMPure XP beads (Beckman Coulter) and eluted in 11 uL of 10 mM UltraPure Tris pH 8. The DNA concentration for each sort was measured and the quality and size of the DNA amplicons was assessed using the Bioanalyzer. The individual samples were combined such that the multiplexed sample included 14 pools distinguished by an index that corresponded to cells collected at the same 1.68 µM LC3B or LC3B LDS\* concentration in each consecutive sort. The multiplexed sample was submitted for sequencing on the Illumina NextSeq500 using 150 nt paired-end reads and custom-designed sequencing primers (ol\_JD814-6). Supplementary file 1 provides an overview of the procedures used to (1) label the individual samples with an index indicating the sort of origin, (2) prepare the library for sequencing, (3) sequence with custom Illumina primers. Index and primer sequences are listed for each sample.

**NGS Data Processing and Analysis**

Demultiplexed sequencing data were merged with bbmerge using maq=20. We used the sequencing data to determine the number of clonal cells (cells displaying the same peptide) that were found in each consecutive sort. We calculated the frequency (of each clone in the collected gate of sort (x) as ci,sort(x)/Ni,sort(x) where ci,sort(x) is the number of raw reads for sequence i in sort (x) and Ni,sort(x) is the total sequencing reads of all clones obtained for sort(x), i.e. Ni,sort(x) = . The enrichment ratio (ER) was then calculated for each peptide to evaluate the enrichment of each clone through the sorting trajectory as follows: where is the frequency of each sequence in the naïve input library (Rubin et al. 2017).

The T7-pep library contains many point mutations and frame shifts. We collapsed hits that we judged to be variants of the same sequence based on sequence clustering using ALFAT-Clust with the pre-set thresholds, selecting the sequence that enriched farthest in the sort as the representative sequence (Chiu and Ong 2022). We then filtered for clones that had a read count of ≥ 10 reads in the input library. Sequences were translated using Biopython (Cock et al. 2009) and mapped to the human proteome using NCBI BLAST. Complete data corresponding to the read counts, for each sequence at every sort, are provided and have been deposited at xxx with accession number xxx. Supplementary File 2 contains the pre-collapsed data and Supplementary File 3 contains the post-collapsed data used in this study. Supplementary File 3 contains the peptide amino-acid sequence, raw counts, ER values, BLAST results, and the z-score. Supplementary File 4 is filtered solely for peptides that enriched through sort 6 and additionally contains AlphaFold pLDDT scores that account for the level of disorder within the hit peptide region.

**LC3B and LC3B LDS\* Scoring Metrics**

With the thresholds mentioned above, 12,158 sequences from the 487,021 of the input library were present in the final sort 6 using LC3B. Among these sequences, clones were further classified as high-enriching binders if the average z-score over sorts 4-6 was greater than or equal to 1.70. The z-score was defined as z = where X is the ER of the given peptide in sorti and and are the mean and standard deviation for the distribution of ER values for all peptides in sorti. The threshold of 1.70 was chosen based on BLI affinity measurements to minimize the number of false positive clones (Figure 2B). Z-scores for each clone that reached sort 6 can be found in Supplemental File 4.

Three scoring metrics were used to assess the behavior of clones in the LC3B LDS\* sort: (1) the ER using clones that enriched through sort 5 (ER5) was calculated across all three LC3B LDS\* enrichment sorts and their trajectories throughout the three-rounds were analyzed manually. (2) ER5 was used to compare peptide enrichment between canonical sort 6 (ER6/ER5) and LDS\* sort 1 (ER LDS\*1/ER5). Promising clones were identified that may bind non-LDS sites by selecting for those that were enriched in ER LDS\*1/ER5 and depleted in ER6/ER5. (3) The ratio of clone frequencies in the binding versus non-binding gate (B/NB) was calculated for each LDS\* sort and canonical sort 6. Clones that continually enriched in the B/NB ratio indicated they were being selected for binding to LC3B LDS\*. Top-scoring clones for each metric were manually selected and measured for binding to LC3B and LC3B LDS\* with BLI. Validated peptide binders scored highly across all three metrics. All LDS\* enrichment data and associated metrics can be found in Supplemental File 5

**Identification of Previously Reported Interaction Partners of LC3B**

To assemble a set of high-confidence interactions, entries for *Homo sapiens* that were experimentally verified to bind one of the hAtg8s were downloaded from LIRCentral (n=115) (Chatzichristofi et al. 2023). Given that these entries could correspond to any of the six hAtg8 paralogs, the LIRCentral referenced primary publications for these entries were manually searched for evidence of experimentally verified binding to LC3B. Of these verified binders, the six that enriched above a z-score of 1.70 are indicated in the LIR row of Figure 2A, Unique *Homo sapien* interactors reported to co-immunoprecipitate with LC3B (MAP1LC3B) (n=550) were downloaded from BioGrid (Stark 2006). Those that enriched in our screen are annotated in the row labeled IP in Figure 2A. To assess whether candidates in our high enriching subset may have biologically relevant interactions with LC3B, gene ontology (GO) terms for all peptides in sort 6 were extracted from UniProt and cross-referenced with those corresponding to LC3B (The UniProt Consortium et al. 2025; Ashburner et al. 2000; The Gene Ontology Consortium et al. 2023). Those that share a GO term with LC3B are indicated in the GO row of Figure 2A.

**PSSM peptide design**

pLogo illustrates the log odds of the significance of over- vs. under-representation of residues in a sample set of sequences relative to a background set (O’Shea et al. 2013). For logo in Figure 3A, clones with a z-score ≥ 1.70 were aligned by the canonical LIR motif [FWY]xx[LVI] and weighted by their z-score (foreground; n = 368) using python and compared to all LIR-containing peptides in the input (background; n = 260755).For logos in Supplemental Figure 6, clones in sort 6 were k-means clustered (n=10) using Clust (Abu-Jamous and Kelly 2018). Peptides in the highest enriching cluster were aligned by the canonical LIR motif [FWY]xx[LVI] (foreground; n = 138) using python and the same process was repeated, using the same background. The process was repeated using subsets of LIR hits from sort 6 based on the identity of the first hydrophobic residue (F, W, or Y) where the background was filtered for the first hydrophobic residue (Supplemental Figure 6). The consensus peptide for each class was defined as the peptide composed of the most over-represented residue in each position and tested for binding to LC3B using BLI.

**Small-scale expression and purification for Biolayer Interferometry**

For BLI experiments, peptides were cloned into a pDW363 biotinylation vector containing SUMO following the BAP and 6xHis tag. Peptides were fused to the C-terminus of SUMO via a GS linker to reduce steric constraints. The peptide constructs were transformed into BL21(DE3) competent cells (New England Biolabs) and expressed at 37 °C at 225 RPM in 20 mL LB with 100 μg/mL ampicillin and a final concentration of 0.05 mM D-(+)-biotin. Cultures were induced with a final concentration of 1 mM IPTG for 4 – 5 hours at 37 °C and then harvested by centrifugation at 5000xg for 15 minutes. Cell pellets were frozen at –80 °C until thawed for purification. Once thawed on ice, cells were lysed with 4 mL/g B-PER reagent (ThermoFisher) and a final concentration of 0.2 mM PMSF. 1 mL of the lysed cells was rotated at 25 °C for 15 minutes before pelleting by centrifugation at 15,000 RPM for 15 minutes. Supernatant was added to 250 μL Ni Sepharose High Performance resin (Cytiva) pre-equilibrated with NiNTA Binding buffer. Resin was washed three times with 1 mL NiNTA binding buffer before elution with 2 mL of NiNTA elution buffer.

**Biolayer Interferometry**

BLI experiments were performed on an Octet Red96 instrument (ForteBio). In vivo biotinylated SUMO-fused peptides, small-scale purified as mentioned above, were immobilized to Octet SA Biosensors (Sartorius) and loaded until a response level of at least 0.3 nm. The ligand-loaded tips were then incubated with increasing protein concentrations and their response measured. Binding experiments took place in 96 well flat bottom polypropylene microplates (Greiner Ref #655209) containing 200 uL of a 50:50 mix of BLI buffer (25 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.5 mM TCEP, 0.1% Tween-20, 1 mg/mL bovine serum albumin (BSA)) and gel filtration buffer lacking glycerol. Each measurement was repeated 2 – 4 times at 25 °C at an agitation speed of 1000 RPM. Dissociation constants for the measured interactions were determined by plotting the equilibrated signal, after subtracting the negative control (immobilized SUMO alone), as a function of protein concentration. The data were then fit using nonlinear regression in GraphPad Prism 10 software.

**Crystallization**

BLM552-571 peptide (DNFDIDDFDDDDDWEDIM) was fused N-terminal to LC3B via a GS linker and cloned into a pGEX vector containing an N-terminal GST and 3C cleavage site. BLM-fused LC3B was expressed in BL21(DE3) competent cells in 2xYT with 100 ug/mL ampicillin. Cells were induced with a final concentration of 1 mM IPTG and expressed shaking at 225 RPM and 25 °C overnight before being harvested by centrifugation at 5000 RPM for 15 minutes at 4 °C. Cell pellets were resuspended in GST binding buffer (140 mM NaCl, 2.7 mM KCl, 20 mM Na2HPO4, 1.8 mM KH2PO4 (1xPBS), pH 7.3) with 0.2 mM PMSF, dounced 20 times, and lysed by sonication. The suspension was centrifuged at 4 °C and 15,000xg for 15 minutes, before ultracentrifugation at 4 °C and 50,000 RPM for 1 hour. Supernatant was then loaded onto a GSTPrep FF 16/10 10 mL column (Cytiva) pre-equilibrated with GST binding buffer, washed, and eluted with GST elution buffer (50 mM Tris, pH 8.0, 150 mM NaCl, 30 mM reduced glutathione). Fractions were analyzed by SDS-PAGE, and those containing the correct molecular weight species were pooled. Cleavage was performed overnight at a ratio of 1:100 3C protease to protein, gently vortexed, and placed on ice at 4 °C. The following morning, the cleavage product was concentrated and purified over a Superdex 75 16/600 column in gel filtration buffer (25 mM Tris-HCl, pH 7.5, 150 mM NaCl). Pure fractions, as assessed by SDS-PAGE, were concentrated and flash-frozen in 50 uL aliquots.

Crystals of BLM-fused LC3B were grown in hanging drops containing 0.1 M HEPES, pH 7.3, 30% w/v PEG 3350, and 0.32 MgCl2. 1 uL of BLM-LC3B was mixed with 1 uL of reservoir solution; crystals grew within 7 days. X-ray diffraction data were collected on a Rigaku Micromax-007 rotating anode with Osmic VariMax-HF mirrors and a Rigaku Saturn 944 detector. Diffraction data were processed with the XDS suite (Kabsch 2010). Phaser (McCoy et al. 2007) was used to solve the structure by molecular replacement using an LC3B structure (PDB 3VTU) (Rogov et al. 2013) as the search model. The molecular replacement solution was refined in PHENIX (Liebschner et al. 2019) with manual fitting in Coot (Emsley et al. 2010) and some improvement of the model geometry with Rosetta. The model was refined to final Rwork/Rfree of 0.216 / 0.262. The X-ray data collection and refinement statistics are summarized in Table SI, and the coordinates have been deposited in the Protein Data Bank (PDB ID: XXXX).

**Encoding**:

We employed three different encoding methods: position-specific scoring matrix (PSSM)[1], onehot[2], and evolutionary scale modeling (ESM-2)[3] to embed peptide sequences into a vector of numbers for subsequent modeling task. Here are the implementation details for each encoding methods. To construct the PSSM, we began by randomly selecting a balanced set of pre-aligned sequences from binders and non-binders. These sequences were then converted into two separate position probability matrices using the alignment\_to\_matrix function from the **Logomaker** Python module [4]: one matrix represented binders (), while the other represented non-binders (). Next, the position probability matrix of non-binders was treated as the background model for calculating the raw PSSM. This was achieved using the transform\_matrix function in **Logomaker**, which transformed the elements in using the background model so that:

To obtain the normalized PSSM, the were normalized on a per-position basis to a range of 0 to 1. This normalization was performed by subtracting the minimum score of each position (the row) from all elements within the same row and then dividing the resulting values by the range of the row:

Special care was taken to handle missing residues, particularly in longer peptide sequences (e.g., 24-mers), where missing residues predominantly appeared in binder sequences. To mitigate the risk of model overfitting caused by these missing residues—thereby preventing an overestimation of model performance on longer sequences—we assigned zero values to the entries corresponding to missing residues in both the raw and normalized PSSM matrices.

For one-hot encoding, we employed the standard approach of mapping each of the 20 amino acids to a unique position in the encoding vector, thereby embedding peptide sequences into a numerical format suitable for downstream analysis. Each amino acid was represented by a binary vector with a length equal to the total number of amino acids, where the position corresponding to the specific amino acid was assigned a value of 1, and all other positions were set to 0. To prevent overfitting on missing residues, we assigned a value of 0 to the positions corresponding to missing residues.

For encoding peptide sequences with ESM2 (version: 2.0.0), we followed the instructions provided by the ESM2 developers. Each peptide sequence was processed iteratively to compute its per-residue representations. Given the relatively small size of our training dataset, we selected the lowest-dimension embedding model, **esm2\_t6\_8M\_UR50D**, which generates a 320-dimensional encoding vector for each residue. These embedding vectors were subsequently used to represent peptide sequences in downstream modeling tasks. However, for Naïve Bayes-based models—namely Gaussian Naïve Bayes (NB), Multinomial Naïve Bayes (mNB), and Complement Naïve Bayes (cNB)—special preprocessing was required. Since both mNB and cNB models only accept non-negative input values, we employed the **MinMaxScaler** from the Python **scikit-learn** module to scale the ESM2 embeddings to the range of 0 to 1. This transformation ensured compatibility with these models while preserving the relative relationships in the original data.

**Modeling method**:

To build classification models and evaluate their performance, we employed five different machine learning algorithms: Random Forest (RF)[5], Naïve Bayes (NB)[6], Multinomial Naïve Bayes (mNB)[7], Complement Naïve Bayes (cNB)[8], and a fully connected neural network (NN)[9]. The dataset was split into ten folds following the method described previously, enabling standard ten-fold cross-validation. For each model-encoding pair combination, we conducted a hyperparameter grid search (details provided below for each model) on the training and validation splits to identify the optimal set of hyperparameters. Using the best-performing hyperparameters, the model was trained and evaluated on the corresponding test subset of the fold. Test accuracy was calculated for each fold, and the median test accuracy across the ten folds was reported to summarize the performance of the model-encoding pair. Additionally, we stratified this evaluation across different peptide lengths to ensure robust performance assessment under varying sequence characteristics. This approach provided a comprehensive evaluation of each method's effectiveness across diverse encoding strategies and peptide lengths.

We utilized the BalancedRandomForestClassifier from the **imbalanced-learn** Python module (version: 0.12.3 [9]) to build the random forest classification model. Unlike the standard one from **scikit-learn**, the BalancedRandomForestClassifier features a built-in sampler that randomly draws a balanced subset from the highly imbalanced training dataset for each tree in the forest. This approach effectively mitigates overfitting to dominant features from non-binder sequences, improving model performance on minority classes. To identify the optimal hyperparameters, we conducted a grid search by varying the following parameters: number of trees in the forest (n\_estimators, from 4 to 50), tree splitting criteria (criteria, either ‘gini’ or ‘entropy’), and cost complexity pruning (ccp\_alpha, from 4e-3 to 1.4e-2) were varied. Bootstrap (False), replacement (True), sampling\_strategy (all) were kept as their default values. To make training process reproducible, we fixed the random state of the model (random\_forest equals to 0).

We employed **GaussianNB**, **MultinomialNB**, and **ComplementNB** from the **scikit-learn** Python module (version\_1.5.2, [10]) to build the Gaussian Naïve Bayes (NB), Multinomial Naïve Bayes (mNB), and Complement Naïve Bayes (cNB) models, respectively. Given the highly imbalanced nature of the training dataset, capturing all relevant features for both binder and non-binder sequences posed a significant challenge for individual models. To address this, we developed a novel approach involving an **ensemble of Bayes estimators**, where each estimator was trained on a balanced subset of the training data. The **RandomUnderSampler** from the **imbalanced-learn** Python module was utilized to randomly downsample non-binder sequences, creating perfectly balanced training subsets for each Bayes estimator in the ensemble. This strategy ensured that the models were exposed to a diverse yet balanced representation of the training data, improving their ability to generalize across both classes. The number of estimators in the ensemble (**n\_estimators**) was treated as a hyperparameter and varied between 10 and 50 for each type of Bayes model (NB, mNB, and cNB). This allowed for optimization of the ensemble size to balance computational efficiency and model performance. This ensemble-based approach, combined with balanced training subsets, was designed to mitigate the challenges posed by the imbalanced dataset and enhance the ability of the models to accurately classify binder and non-binder sequences.

A two-hidden-layer neural network model was constructed using **TensorFlow's Keras API (version: 3.7.0 [11]).** The **ReLU** activation function was applied to the neurons in both hidden layers to introduce nonlinearity. To prevent overfitting, two regularization strategies were employed: an **L2 regularization penalty** with a default value of 0.01 and a **dropout rate** of 0.5 applied to each hidden layer. Additionally, **batch normalization layers** were included before the second hidden layer and the output layer to accelerate training and improve stability. The output layer utilized a **sigmoid activation function**, which produced a predicted probability for the binary classification task. The following hyperparameters were varied to identify the optimal model configuration: batch size (batch\_size, chosen from [64, 128, 256, 512, 1024]), learning rate (learning\_rate, chosen from [1e-4, 1e-3, 1e-2), and hidden layer size (hidden\_unit, chosen from [32, 64, 128, 256]) were varied. Each model was trained for **1,000 epochs**, and the best model was selected based on its **F1 score** on the validation set, balancing precision and recall. The inherent imbalance in the training dataset presented a challenge, as the model could easily be overfit to the dominant features of non-binder sequences. To mitigate this, a custom data generator was implemented with the following features:

1. For every batch, a perfectly balanced subset of the training data was created by randomly sampling binder and non-binder sequences based on the batch size.
2. Over the course of an epoch, all binder sequences were utilized at least once. Over N epochs, all non-binder sequences were also used at least once to ensure full utilization of the training dataset while maintaining a balanced training paradigm.

This generator ensured that the model was trained on representative, balanced subsets of the data in each batch, while still leveraging the full diversity of the dataset across multiple epochs. This approach, combined with regularization techniques, minimized overfitting and improved the generalizability of the model.

**Figures**

**A screenshot of a video game

AI-generated content may be incorrect.**

**Figure 1. Human peptidome library enrichment screening reveals known and novel binders.**

1. Schematic of FACS-based bacterial display enrichment screening. 36-residue peptides (orange) from the human peptidome were expressed as fusions to eCPX, with a FLAG tag, on the surface of *E. coli*. Expression was detected using anti-FLAG APC fluorescence. Binding to LC3B (grey), which was tetramerized through binding to streptavidin conjugated to phycoerythrin (SA-PE, pink), was detected using PE fluorescence. Peptide-expressing cells were sorted based on APC and PE fluorescence and sequenced via NGS. Schematic made using BioRender.
2. Unique sequences detected (n=through five rounds of enrichment sorting as described in the methods. Sort 2 was a negative sort used to eliminate peptides nonspecifically bound to SA-PE and was not sequenced.
3. Peptide enrichment profiles across enrichment sorts. Black lines show trajectories for a random 0.1% of all peptides that persisted to sort 6. Overlaid are the enrichment profiles for the best-enriching peptide (BLM 552-570\*), positive controls from the sort (FYCO11277-1312 and ATG4A363-398), and the worst-enriching peptide that reached sort 6 (EXTL3871-906) colored green, pink, yellow, and blue, respectively.
4. Cumulative density function plot of the average z-scores across sorts 4, 5, and 6. Average z-scores for peptides colored in black, and for peptides from BLM552-570\*, FYCO11277-1312, and ATG4A363-398 colored as in panel C, surpass the threshold of 1.70.
5. Peptide enrichment profiles for the 427 peptides that passed the z-score threshold (black), with peptides from BLM552-570\*, FYCO11277-1312, and ATG4A363-398 in color.

**A screenshot of a computer

AI-generated content may be incorrect.Figure 2. Highly enriching peptides bind LC3B with low micromolar affinity**

1. Screen validation. For peptides with average z-score1.70, the figure shows prior support for the corresponding protein interacting with LC3B. LIR - experimentally-validated LIR motifs annotated in LIRCentral (Chatzichristofi et al. 2023). IP - proteins pulled down with LC3B, from BioGrid. GO – peptides with GO-annotations shared with LC3B, from Uniprot (The UniProt Consortium et al. 2025; Ashburner et al. 2000; The Gene Ontology Consortium et al. 2023).
2. Binding curves and affinities determined by BLI. HEAT3, ATG4A, SCYL1 have previously been immunoprecipitated with LC3B whereas BLM has not.
3. Relationship between average z-score over rounds 4-6 and affinity for monomeric LC3B as determined by bio-layer interferometry (BLI). Colors indicate peptides as in Figure 1C. Shapes indicate five classes of peptides: peptides with LIR motifs validated to bind to LC3B (circles), peptides corresponding to proteins pulled down with LC3B (triangles), peptides from proteins with shared GO terms with LC3B (diamonds), peptides meeting both IP and GO criteria (squares), and unannotated peptides (stars). KD values represent the mean over at least two replicate experiments.

**A close-up of a computer screen

AI-generated content may be incorrect.Figure 3. Enriched LIR-containing peptides show a preference for acidic flanking sequence and a W-type core LIR.**

1. pLogo (O’Shea et al. 2013) showing over- and under-represented residues in LIR-containing peptides with z-score 1.70, relative to all LIR-containing peptides in the input library. The full-length logo is in Supplemental Figure 5.
2. BLI binding curves for LC3B binding to peptides pCONSLIR (EEEVEEKEEEDDDEEWEILDIEEGSDSEQKLISE), Ank21578-1613  (VQSSRSERGLVEEEWVIVSDEEIEEARQKAPLEITE), and the 36-mer of FYCO11277-1312 identified by the sort (DAVFDIITDEELCQIQESGSSLPETPTETDSLDPNA). Error reported as the standard deviation of two or more technical replicates.
3. Sequential truncations of pCONSw measured via BLI (mean ± sem, *n* ≥ 3). Statistical analysis using one-way analysis of variance (ANOVA) test; mean ± s.d.; \*\*\*\**p* ≤ 0.0001; \*\*\**p* ≤ 0.001; \*\**p* ≤ 0.01; \**p* ≤ 0.05; ns not significant.
4. Structure of BLM552-571 bound to LC3B, determined at 2.2 resolution. Here, and in panels e-f, the BLM peptide is green and the LC3B surface is colored by hydrophobicity (ChimeraX). Hydrophobic pockets HP1 and HP2, and the N-terminal flanking residues, are indicated.
5. N-terminal acidic residues D564-6 of BLM interact with R10, R11, and K51 of LC3B. Contacts within 4 are shown as black dashed lines, and those greater than 4 are colored in dark gray, with the distances labeled.
6. In HP1, W567 contacts LC3B I23, P32, K51, L53, and F108.
7. In HP2, E568 contacts R70 and D569 (depicted as black dashed lines). LC3B residues within 4 of BLM I570 are shown in sticks. BLM M571 contacts LC3B P55 and I66.

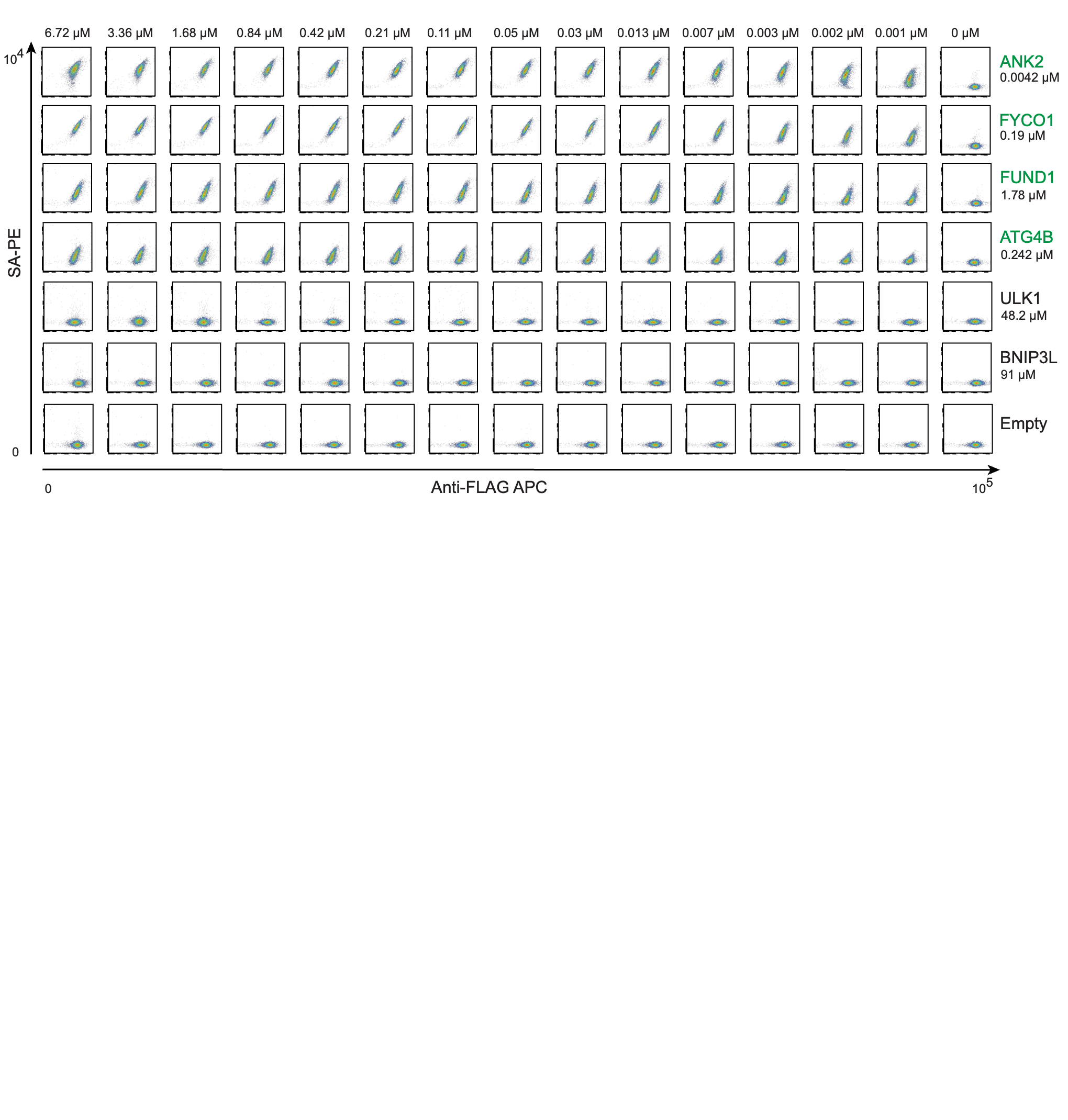
A close-up of a computer screen

AI-generated content may be incorrect.

**Figure 4. LIR+ binding motifs and LC3B LDS\* enrichment sorting**

1. Affinities of LC3B for peptides from PAR118-37, CTSL2233-262, and DYH12422-445\* (mean ± s.e.m, *n*≥3). Candidate binding sites that were mutated to Ala are red in the mutated sequence. Nonbinding peptides, determined to have no binding response up to 40 µM LC3B, are marked in light gray and labeled with N.B.
2. Structure of LC3B (PDB 3VTU) (Rogov et al. 2013) showing Phe52 in HP2 and Leu53 in HP1, which were mutated to Ala in LC3B LDS\*. Known phosphorylation sites near the LDS are shown in black (T12, T50) (Nieto-Torres et al. 2021; Shrestha et al. 2020; Cherra et al. 2010).
3. Enrichment profiles for nine peptides through the three rounds of the LC3B LDS\* enrichment sort. The four peptides that bound with measurable affinity to LC3B LDS\* (colored in shades of blue) show enrichment through the three rounds. Known, canonical LC3B LDS binders (colored in purple and red) show depletion throughout the LDS\* sort and are shown for comparison. Black lines (alpha 0.1) indicate a random sampling of 10% of peptides that persisted through LC3B LDS\* sort 3.
4. Affinities of LC3B (blue) and LC3B LDS\* (light red) for peptides from FYCO11277-1312 and ATG4A363-398 and for four confirmed hits from LC3B LDS\* sorting, PPM1H436-471, OSBL71-36, PEP2, TRIM5288-120. (mean  ± s.e.m, *n*≥3). Binding profiles for FYCO11277-1312 and ATG4A363-398 are shown in Supplemental Figure 11.

**Supplemental Figures**



**Figure S1. Dynamic range of well-studied LIR motifs binding to pre-tetramerized LC3B in bacterial display**

Positive control peptides Ank21578-1613, FYCO11264-1299, FUND12-37, ATG4B372-407, ULK1341-376, BNIP3L20-55, and negative control Empty (eCPX-FLAG with no peptide) were tested for binding to serially diluted pre-tetramerized LC3B using bacterial display. Peptides colored in green show concentration-dependent, saturable binding. Reported affinities were obtained from the corresponding reference entry in LIRCentral (Chatzichristofi et al. 2023). Results from this analysis led to the selection of 1.68 µM LC3B as the optimal sorting concentration.

**A diagram of different types of objects

Description automatically generated with medium confidence**

**Figure S2. Canonical enrichment sort with pre-tetramerized LC3B**

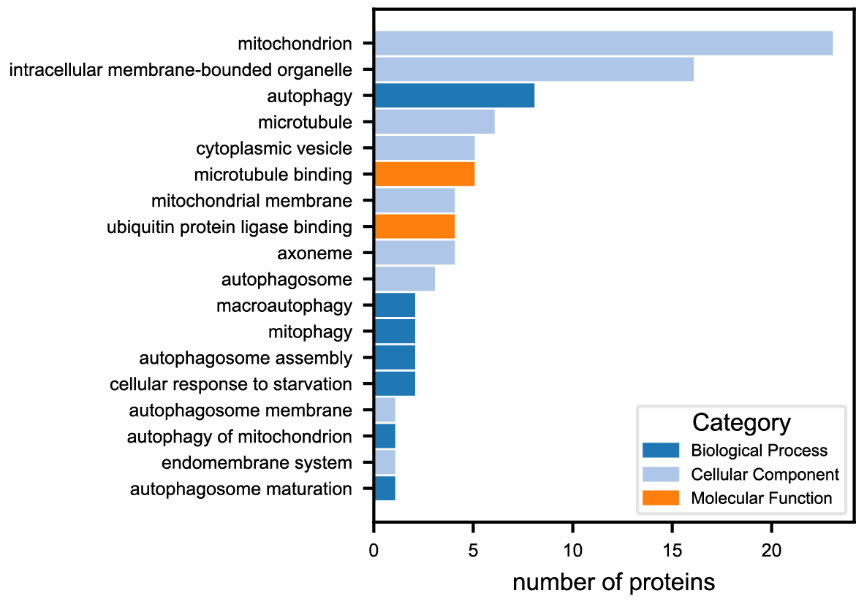
The input library was sorted for six rounds, selecting for binding to streptavidin-tetramerized LC3B. Positive controls FYCO11264-1299 and ATG4B372-407 are shown for sort 1, as is negative control Empty (eCPX-FLAG with no peptide). Sort 2 removed clones that bound to SA-PE. The remaining four sorts collected the indicated percentage of LC3B binding cells using the indicated gate.

**A screen shot of a cellphone

Description automatically generated**

**Figure S3. Behavior of known LC3B-binding peptides in the sorting experiment.**

Heatmaps show the extent to which LIRCentral-annotated peptides contained in the input library (rows), were observed in each round of sorting. Hashing indicates the peptide was not identified in that sort. **(a)** High-affinity binders. The peptide identified from the sort is shown with the core LIR in bold and any mutations from the human sequence in red. The peptide tested for binding to LC3B is shown on the right. Previously measured affinities are as follows: FYCO11273-1297 (0.19 µM ITC) (Cheng et al. 2016) , ATG4B358-394 (0.47 µM BLI), TCPR21403-1411 (0.67 µM ITC) (Stadel et al. 2015), NEDD4256-278 (0.94 µM FP) (Qiu et al. 2017), FUND110-25 (1.78 µM ITC) (Lv et al. 2017), KBTB7658-674 (2.3 µM ITC) (Genau et al. 2015), NBR1722-739 (2.9 µM ITC) (Rozenknop et al. 2011), PKHM1629-642 (6.3 µM ITC) (Rogov et al. 2017),SEC62352-370 (25.9 µM BLI) (Fumagalli et al. 2016), ULK1349-369 (48.2 µM BLI) (Wirth et al. 2019). **(b)** Peptides annotated as non-binding in LIRCentral (Chatzichristofi et al. 2023). Protein residues that define the peptide present in the sort are indicated, based on the Uniprot (The UniProt Consortium et al. 2025) entry.

****

**Figure S4. Gene Ontology (GO) terms shared with LC3B**

GO terms shared between proteins with z-score ≥ 1.70 and LC3B, colored by GO ontology category. Shared GO terms were sourced and processed from UniProt (The UniProt Consortium et al. 2025) (see Methods).

A blue circle with black text

AI-generated content may be incorrect.

**Figure S5.** **LIR content of the top-enriching hits.**

The pie chart includes peptides from the screen with z-score ≥ 1.70. Canonical LIR motifs are defined as [FWY]xx[LVI] and LIR-adjacent motifs are defined as [FWYLVI]xx[FWYLVI]. For this chart, the LIR-adjacent category excludes peptides in the canonical LIR category. HPQ indicates peptides containing the biotin mimic “HPQ” motif, and Novel designates peptides without any recognizable regular expression pattern.

**A screenshot of a computer

AI-generated content may be incorrect.**

**Figure S6. pLogos composed of peptides with F-type, W-type, and Y-type LIRs**

pLogos showing the enrichment or depletion of residues in sequences that match different LIR motifs, relative to a background of all LIR-containing peptides in the input library. (a) peptides with average z-score ≥ 1.7 that match the canonical LIR motif. (b-d) peptides that match the canonical LIR and contain either F, W, or Y in the first hydrophobic position (top to bottom). Acidic resides are colored dark red, and residues with statistically significant enrichment as calculated by pLogo are shown in bright red. Core LIR motif positions are highlighted in yellow, with [FWY] and [LVI] in black.

**A diagram of a protein

AI-generated content may be incorrect.**

**Figure S7. Known c-terminal extension glutamates contacting LC3B R70**

Known peptides containing C-terminal extensions engage LC3B R70. These peptides include FYCO1 (colored in cornflower blue), ANK2 (colored in salmon) , and ANK3 (colored in forest green) (PDB: 5D94 Olsvik et al. 2015; PDB: 5YIQ, 5YIS Li et al. 2018a; PDB: 5CX3 Cheng et al. 2016). The residues making the interaction are E1283 of FYCO1, E131 of ANK2, and E131 of ANK3. The complexes were aligned using the best aligning pair of chains between reference and match in ChimeraX, with only the LC3B of 5D94 shown for visual clarity. Hydrogen bonds between FYCO1 and LC3B are shown in black dashes.

**Figure S8. Model predictions compared to validated BLI hits. [WAITING FOR FINAL MODEL]**

**A colorful letters on a black background

AI-generated content may be incorrect.**

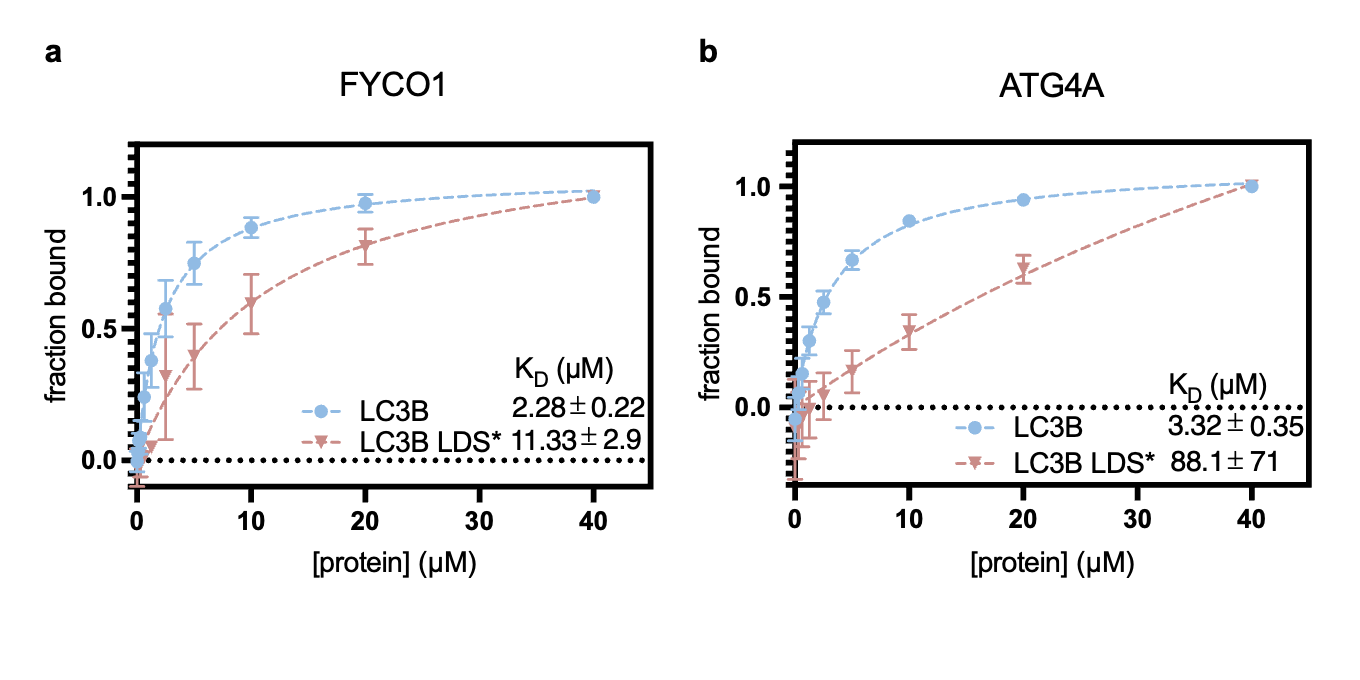
**Figure S9. HPQ False Positive peptides**

Identification of HPQ by XSTREME (Grant and Bailey 2021) as an enriched binding motif in the sort. The highest enriching peptides (n = 427) were input into XSTREME with the input library as background. HPQ was the third highest enriching motif.

**A diagram of different types of blotches

Description automatically generatedFigure S10. LC3B LDS\* enrichment sort**

An additional three rounds of bacterial display sorting were performed for binding to pre-tetramerized LC3B LDS\*, using the population collected from round 5 of the LC3B sorting experiment as input. Signals for positive controls FYCO11264-1299 and ATG4A363-398 and the empty-vector control are shown for 1.68 µM LC3B (blue) overlayed with 1.68 µM LC3B LDS\* (red). The binding (upper right gate) and nonbinding (lower right gate) populations for each LC3B LDS\* sort, and a sixth round of LC3B binders, were collected and sequenced. The percentage of cells collected is indicated in each gate.



**Figure S11. FYCO1 and ATG4A binding to LC3B and LC3B LDS\***

(a) FYCO11264-1299 and (b) ATG4A363-398 binding to monomeric LC3B (blue) and monomeric LC3B LDS\* (pink) via BLI. Error reported as the standard deviation of two or more technical replicates.

**Table 1. The affinities of LIR-containing peptides for LC3B, determined using BLI.**

Binding data for other peptides can be found in Supplementary Table 1.

