Computational Modeling of C-Terminal Tails to Predict the Calcium-Dependent Secretion of Endoplasmic Reticulum Resident Proteins

Kathleen A. Trychta[1†], Bing Xie[2†], Ravi Kumar Verma[2], Min Xu[2], Lei Shi[2‡]* and Brandon K. Harvey[1‡]*

[1]Molecular Mechanisms of Cellular Stress and Inflammation Unit, National Institute on Drug Abuse, National Institutes of Health, Baltimore, MD, United States, [2]Computational Chemistry and Molecular Biophysics Section, National Institute on Drug Abuse, National Institutes of Health, Baltimore, MD, United States

The lumen of the endoplasmic reticulum (ER) has resident proteins that are critical to perform the various tasks of the ER such as protein maturation and lipid metabolism. These ER resident proteins typically have a carboxy-terminal ER retention/retrieval sequence (ERS). The canonical ERS that promotes ER retrieval is Lys-Asp-Glu-Leu (KDEL) and when an ER resident protein moves from the ER to the Golgi, KDEL receptors (KDELRs) in the Golgi recognize the ERS and return the protein to the ER lumen. Depletion of ER calcium leads to the mass departure of ER resident proteins in a process termed exodosis, which is regulated by KDELRs. Here, by combining computational prediction with machine learning-based models and experimental validation, we identify carboxy tail sequences of ER resident proteins divergent from the canonical "KDEL" ERS. Using molecular modeling and simulations, we demonstrated that two representative non-canonical ERS can stably bind to the KDELR. Collectively, we developed a method to predict whether a carboxy-terminal sequence acts as a putative ERS that would undergo secretion in response to ER calcium depletion and interacts with the KDELRs. The interaction between the ERS and the KDELR extends beyond the final four carboxy terminal residues of the ERS. Identification of proteins that undergo exodosis will further our understanding of changes in ER proteostasis under physiological and pathological conditions where ER calcium is depleted.

Keywords: ER calcium, exodosis, KDEL receptor, endoplasmic reticulum, ER retention sequence

INTRODUCTION

The classical secretory pathway involves the movement of proteins from the endoplasmic reticulum (ER) to the Golgi and then to the cell surface. Resident proteins of the ER counteract this flow by using short peptide sequences. For soluble proteins, a carboxy-terminal (C-terminal) ER retention/retrieval sequence (ERS) interacts with KDEL receptors in the Golgi membrane to return proteins back to the ER lumen in a COPI-mediated manner (Munro and Pelham, 1987; Lewis and Pelham, 1992a; Orci et al., 1997). Mammals express three KDEL receptor isoforms, KDELR1, KDELR2, and

KDELR3, all of which function as part of a Golgi-to-ER retrieval pathway to maintain the ER proteome (Lewis and Pelham, 1990; Lewis and Pelham, 1992b; Hsu et al., 1992; Collins et al., 2004; Raykhel et al., 2007). Maintaining ERS-containing proteins within the ER lumen is necessary for these proteins to carry out the diverse functions of the ER including protein trafficking and modification and lipid and carbohydrate metabolism. Loss of these proteins from the ER lumen may compromise ER function and lead to gain of function elsewhere inside or outside of the cell.

In mammals, the KDEL (lysine-aspartate-glutamate-leucine) motif is considered the canonical ERS. The first observations of an ER retention sequence noted that three ER resident proteins (BiP/Grp78, Grp94, protein disulfide isomerase) shared a C-terminal KDEL sequence and that deletion of this sequence resulted in secretion of the protein (Munro and Pelham, 1987). Further, appending this KDEL sequence to a secretory protein resulted in a failure of secretion and protein accumulation within the ER (Munro and Pelham, 1987). Subsequent studies identified KEEL, RDEL, and QEDL as putative ERS (Fliegel et al., 1990; Mazzarella et al., 1990; Mazzarella et al., 1994). A bioinformatic approach intended to identify ER localization motifs found 46 putative ERS even while using a data set restricted to human proteins with the terminal four amino acids *XX*(DE)(FLM) (Raykhel et al., 2007). As the name implies, the current understanding of ER retention sequences focuses on their ability to retain proteins in the ER. However, recent findings suggest that an important functional significance of ERS lies in their ability to regulate protein secretion. While Raykhel et al. has already provided an analysis of C-terminal motifs that localize proteins to the ER, no analysis has yet been reported that systematically investigates the connection between C-terminal motifs and secretion. This definition of what constitutes an ERS was further expanded upon by Alanen et al. showing that the -5 and -6 amino acids also play a role in determining the ER localization of a protein (Alanen et al., 2011). In fact, expanding to five or six residues for some protein tails caused ER localization whereas the four amino acids did not (Alanen et al., 2011). Thus, the amino acids of the carboxy terminus that determine ER localization can extend beyond the final four C-terminal residues of the protein.

ERS play a role in regulating the secretion of proteins as well. A study of four ER luminal proteins with a C-terminal KDEL (BiP/Grp78, Grp94, protein disulfide isomerase) showed that these proteins were secreted from the cell following treatment with calcium ionophores suggesting that ERS-containing proteins are sensitive to changes in intracellular calcium (Booth and Koch, 1989). More recent studies on ER resident esterases and the ER resident protein mesencephalic astrocyte-derived neurotrophic factor (MANF) demonstrated that the secretion of these proteins is increased specifically following ER calcium depletion (Henderson et al., 2013; Glembotski et al., 2012; Trychta et al., 2018b). Furthermore, in several diseases ERS-containing proteins have been found in extracellular fluid (e.g., serum, plasma, saliva, cerebrospinal fluid) and their presence may be a symptom of disease (**Supplementary Table S1**). In fact, ER calcium depletion triggered a mass redistribution of ERS proteins from the ER lumen to the extracellular space, a process termed ER exodosis, which is dependent on the level of KDEL receptors (Trychta et al., 2018a). Other triggers of ER exodosis, such as ischemia, have also been identified (Trychta et al., 2018a). A comparison of >100 putative ERS appended to the C-terminus of *Gaussia* luciferase (GLuc) protein revealed varied responses to ER calcium depletion with some ERS being more readily secreted in response to ER calcium depletion than others (Trychta et al., 2018a). This raises the possibility that ERS-containing proteins could act as biomarkers of exodosis in a variety of disease states. The ability to predict whether an ERS confers secretion following ER calcium depletion would be useful for understanding how pathological states affect cellular proteostasis and how mutations may affect trafficking of ERS-containing proteins.

METHODS

Plasmid Construction
GLuc-ERS proteins were created as previously described resulting in pLenti6.3-MANF sigpep-GLuc-ERS constructs (Trychta et al., 2018a). Briefly, custom forward and reverse oligonucleotides coding for each specific seven amino C-terminal tails were synthesized (Integrated DNA Technologies), formed into an oligonucleotide duplex, then inserted into a pLenti6.3-MANF sigpep-GLuc vector by ligation (Ligate-IT, Affymetrix). Immediately following ligation, DNA was transformed and propagated in NEB stable competent cells cultured at 30°C (New England BioLabs). The resulting DNA plasmids encoding each ERS tail fused to *Gaussia* luciferase with a MANF signal peptide were sequence verified for proper ligation of the tail.

Cell Culture and Transfections
Human SH-SY5Y neuroblastoma cells were maintained in growth media composed of Dulbecco's Modified Eagle Medium (DMEM 1X + GlutaMAX, 4.5 g/L D-glucose, 110 mg/L sodium pyruvate; Invitrogen) containing 10% bovine growth serum (BGS; GE Life Sciences), 10 units/mL penicillin, and 10 μg/ml streptomycin. Cells were grown at 37°C with 5.5% CO_2 in a humidified incubator. For reverse transfections, 5×10^4 SH-SY5Y cells in 90 μl of growth media containing no antibiotics were plated in opaque TC-treated 96-well plates containing Xfect (Clontech)/DNA transfection complexes. For each well, the transfection complexes contained 200 ng plasmid DNA of unique GLuc-ERS tail and 0.06 μl of Xfect in a final volume of 10 μl. After 28–29 h, a full media exchange into growth media containing 1.5% BGS was performed. Thapsigargin (Tg; Sigma) treatments began 16–17 h after the media exchange and lasted for 8 h, a time point where we have demonstrated ERS proteins are secreted in response to thapsigargin (Trychta et al., 2018a). At the end of each experiment, media was removed from the wells and cells were lyzed directly in the plate and used for luciferase assay. Cells were rinsed with 1x PBS and lyzed with 75 μl of lysis buffer [50 mM Tris (pH 7.5), 150 mM NaCl, 1% NP40, 1x protease inhibitors]. For all experiments, vehicle controls were used at a concentration equivalent to the Tg treatment.

Gaussia Luciferase Secretion Assay

Assays for *Gaussia* luciferase activity (GLuc) were performed as previously described (Henderson et al., 2014; Henderson et al., 2015). The GLuc substrate was PBS containing 10 µM coelenterazine (Regis Technologies). Coelenterazine stock solutions were prepared at 20 mM in acidified methanol (10 µl of 10 N HCl per 1 ml of methanol) and stored at −80°C as single use aliquots. For luciferase secretion assays, 5 µl of cell culture medium from each well was transferred to a new opaque walled plate. Luciferase levels were determined using a plate reader with an injector setup (BioTek Synergy II) that allowed samples to be read directly after substrate injection. 100 µL of substrate was injected to the well containing cell culture medium or lysates. For lysate readings, 100 µl of GLuc substrate was injected directly into the tissue culture plate already containing 75 µl lysis buffer. Samples were read at 25°C with a sensitivity of 100 and a 0.5 s integration time with a 5 s delay following coelenterazine substrate injection. Data are presented as a "secretion index" representing the ratio of extracellular to intracellular GLuc where a higher number indicates greater secretion. For each experiment the secretion index of each sample was normalized to the average secretion index of the vehicle treated samples before combining the data across three independent experiments.

KDELR2 Overexpression

Lentivirus encoding Myc-FLAG tagged KDELR2 was previously described (Henderson et al., 2013). Lentiviral vectors expressing KDELR2 or a control empty vector were titered using the Lenti-X p24 rapid titer kit (Takara). 8×10^5 SH-SY5Y in 2 ml growth media were plated in a 6-well plate. 24 h after plating cells were transduced with LV-KDELR2 or LV-Empty (multiplicity of infection = 2). After a 48 h incubation, cells were re-plated for reverse transfections as described above.

Machine Learning Based Prediction Model

We built recurrent neural network (RNN) models to predict Tg-induced stress response from the amino-acid sequences of last seven residues of proteins (C-terminus tails), using the TensorFlow framework (version 1.7). Each seven-residue sequence was encoded by a 7 by 20 matrix using the one-hot method, i.e., for each residue, only one element corresponding to the residue type was set to be 1 and the rest were 0 with the entire feature length of 20 representing the possible natural amino acids.

The training data was obtained from previously published results (Trychta et al., 2018a), which include 95 seven-residue tail sequences (the first batch). The cost function in the modeling process was the mean-squared error between the predicted and the true values. Based on the optimization of this function using the training data set, we chose the following setup for our models. Long Short-Term Memory network (LSTM) was chosen for our task, so that the long-term dependencies in the sequences can be captured. Only one layer of LSTM was used and the size of the hidden state of an LSTM unit was set to be 64. The LSTM layer was connected to a fully connected layer, followed by one single rectified linear unit (RELU) node to produce the output. We used the Adam algorithm to minimize the cost function with the learning rate of 0.001. The hyperparameters here were decided by

10-fold cross validation. To avoid overfitting, the network was trained with a drop-out rate 0.2, i.e., output from the 20% of the randomly chosen neurons was ignored, with 3,000 iterations and during the prediction all the parameters were used.

To eliminate the stochastic element of the RNN prediction, we built multiple models to evaluate how many models were needed for the estimations to converge. Thus, for each of 64 independently built RNN models, we first calculated the Pearson correlation coefficient between the prediction and the experimental data (**Supplementary Figure S1A**). Then for each number of models, we carried out random selection of the models from this pool (one hundred times bootstrapping) and calculated the average and the standard deviation of the Pearson R. Our results show that the slope of the standard error curve is <0.0003 when using more than 15 models, which we define as convergence (**Supplementary Figure S1B**). Therefore, we chose to use the average of the predictions from 32 models in our study.

Assembling a Validation Tail Batch From Artificial Sequences and Human Proteome
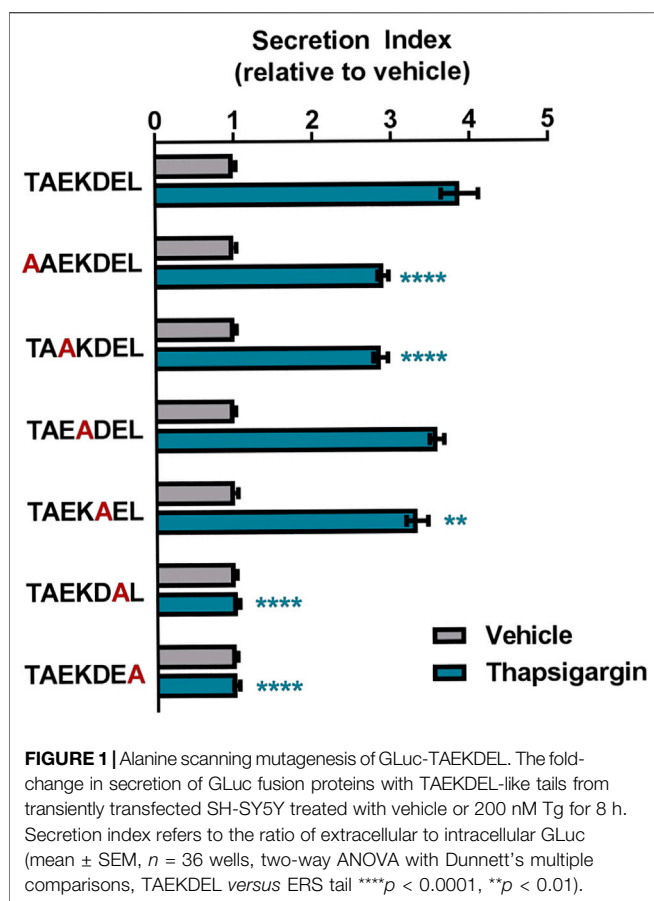
We obtained the high-frequency residue types for each position of the first batch 95 tail sequences and generated all permutations. A total of 463,736 sequences were generated and then sorted by their predicted Tg-induced stress response using our RNN models.

The entire human proteome (GRCh38.p12) consisting of 113,620 sequences was downloaded from NCBI (http://www.ncbi.nlm.nih.gov). We updated this set against the daily updating RefSeq (until 02/28/2019). For each entry in the resulting dataset, the C-terminal seven residues were extracted, and a final set of filtering was performed by removing sequences containing unnatural residues. A final set of 112,477 human protein C-terminus tails were retained, which included 31,872 unique C-terminus tails. The Tg-induced stress response prediction was then carried out for each tail using our RNN models.

We then heuristically selected 104 C-terminal seven-residue sequences, based on the following considerations: the predicted strength of responses, the novelty of the corresponding human proteins compared to the first batch, and the synthetic feasibility. The resulting batch (second batch) includes 48 artificial sequences and 56 human sequences with 65 sequences predicted to have strong responses and 39 to have weak responses.

WebLogo and Similarity Scores

The sequence logos were created using WebLogo3.6.0 (http://weblogo.threeplusone.com) to graphically represent the probability of an amino acid residue type being present at each position by the height of each symbol. The similarity scores were calculated by an in-house script using the Point Accepted Mutation (PAM) 250 matrix. For each pair of aligned residues, the substitution scores were looked up from the matrix; then the similarity score for each pair of compared peptides were obtained by adding the substitution scores of -1 to -7 positions together.

FIGURE 1 | Alanine scanning mutagenesis of GLuc-TAEKDEL. The fold-change in secretion of GLuc fusion proteins with TAEKDEL-like tails from transiently transfected SH-SY5Y treated with vehicle or 200 nM Tg for 8 h. Secretion index refers to the ratio of extracellular to intracellular GLuc (mean ± SEM, n = 36 wells, two-way ANOVA with Dunnett's multiple comparisons, TAEKDEL versus ERS tail ****p < 0.0001, **p < 0.01).

Molecular Docking

The selected tail sequences were first drawn with Chemdraw (version 19.1), and then constructed with the Ligprep module of the Schrodinger (version 2019-4) suit using the OPLS3e forcefield (Roos et al., 2019). The "AEKDEL" of "TAEKDEL" was obtained from the KDELR2 crystal structure (PDB ID 6I6H), and we added "T" to the N-terminus of the peptide. The original KDELR2 structure (PDB ID 6I6H) was first processed by the protein preparation wizard module of the Schrodinger suit, and was then minimized with the OPLS3e forcefield.

We performed docking for the selected peptides with the extended sampling of the induced fit protocol (Sherman et al., 2006) implemented in Schrodinger Suite. The docking box center was determined by the center of mass of the peptide AEKDEL bound in the KDELR2 structure, and the box size was set to be 25 Å. By refining the residues within 5.0 Å of the ligand poses, the peptides were docked into the prepared KDELR2 model with the last residue "L" restrained at the original position found in the crystal structure.

Molecular Dynamic Simulations

With the protein prepared as above, we also kept the crystal water inside the protein and protonated HIS 12 to its positively charged form known as "HIP" (Wu et al., 2020). The POPC lipid and water was added through Desmond setup. We kept the system charge as neutral by adding ions and the salt concentration of the

system was 150 mM. The prepared system contained around 77,000 atoms, which includes ~14,000 water molecules and 221 lipid molecules. For the peptides LIGSLEL and CIHSPDL, we chose the best docking pose as the starting position of the MD simulations. For TAEKDEL, we adopted the pose found in the crystal structure as the starting position.

The MD simulations were carried out with Desmond Molecular Dynamics System (version 6.1; D.E. Shaw Research, New York, NY) using the OPLS3 force field and SPC water model. The system was first relaxed with Brownie Dynamic NVT condition with all solute heavy (non-hydrogen) atoms restrained for 100 ps, and then the systems were equilibrated with NPT for 30 ns. After the equilibration, we collected two 1,200 ns MD trajectories for each receptor-peptide complex system. The backbone of the transmembrane helices was restrained throughout the simulations.

RESULTS

Single Amino Acid Residue Changes Affect ER Retention/Retrieval Sequence Function

In order to examine how changes in C-terminal ERS affect protein secretion in response to ER calcium depletion, we used a similar approach to Raykhel et al. and created a set of GLuc constructs that were identical except for their seven-residue ERS (Raykhel et al., 2007; Alanen et al., 2011). Using thapsigargin (Tg), a selective inhibitor of the sarcoplasmic endoplasmic reticulum ATPase (SERCA), we depleted ER calcium and showed that C-terminal sequences beyond the canonical KDEL motif conferred secretion in response to ER calcium depletion (Thastrup et al., 1990; Trychta et al., 2018a). To explore which permutations within the canonical KDEL motif affect the ER exodosis phenotype, we used TAEKDEL, the ERS of endoplasmin/Grp94 and BiP/Grp78, as a template to carry out an alanine scanning mutagenesis study. Our results revealed that an alanine substitution in the -1 or -2 position drastically reduced Tg-induced secretion when compared to TAEKDEL, substitutions in the -5 and -7 positions also reduced the secretion, compared to that of TAEKDEL (**Figure 1**). We observed that an alanine substitution in the -1 or -2 position led to a high level of GLuc secretion under vehicle conditions such that no further enhancement of secretion could be seen following thapsigargin treatment (**Supplementary Figure S2**). Alanen et al. found that the residues upstream of the C-terminal KDEL were important for determining ER localization of proteins and we show that the -5 and -7 positions play a role in determining whether a protein is secreted in response to ER calcium depletion (Alanen et al., 2011). These data along with the observations of Alanen et al. support that ERS may play a role in regulating both protein retention (ER localization) and protein secretion (Alanen et al., 2011). We show that changing a single residue within an ERS has the potential to change ER resident protein secretion caused by ER calcium depletion. The discovery that the amino acid composition of an ERS can affect its ability to undergo ER exodosis has ramifications for understanding how ER proteostasis changes in times of cellular stress.
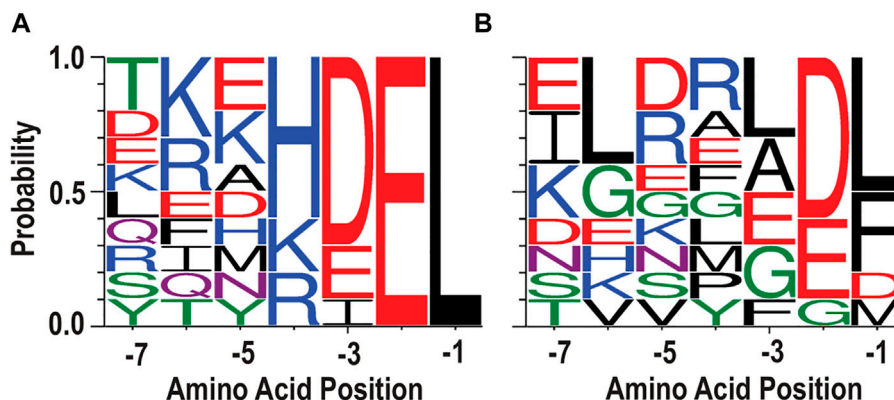
FIGURE 2 | Sequences divergent from KDEL respond to ER calcium depletion. **(A)** WebLogo depicting ten seven amino acid carboxy-terminal peptide sequences that conferred the highest secretion in response to Tg. **(B)** WebLogo depicting the ten carboxy terminal peptide sequences with lowest Tg response. For WebLogo settings, we used the parameters of data type "-D" as FASTA, sequence type "-A" as protein, color scheme as "chemistry", and unit as "probability".

Non-Canonical ERS Undergo Exodosis

Previously, we evaluated 95 different C-terminal tails from human proteins in a cell culture model of ER exodosis. We attached each seven-residue tail to the C-terminus of GLuc and monitored Tg-induced secretion responses, and found that "KKKKDEL" represented the most dominant residues at each position in the C-termini (Trychta et al., 2018a). However, when parsing the data to look at the tail sequences that result in the greatest Tg-induced secretion (top 10, **Figure 2A**), it is difficult to identify a consensus sequence motif that would predict secretion responses, i.e., the tails resulting in a strong response do not have a clearly dominant residue type (or residue set with similar physicochemical properties) at each position that would clearly differentiate them from the tails resulting in low responses (bottom 10, **Figure 2B**). For example, the negative charged residue Glu appears in -2 and -3 positions in both **Figures 2A,B**. In addition, while Leu is the only residue type appearing in the -1 position among the top 10 tail sequences, it is also the most probable residue in the bottom 10 tail sequences.
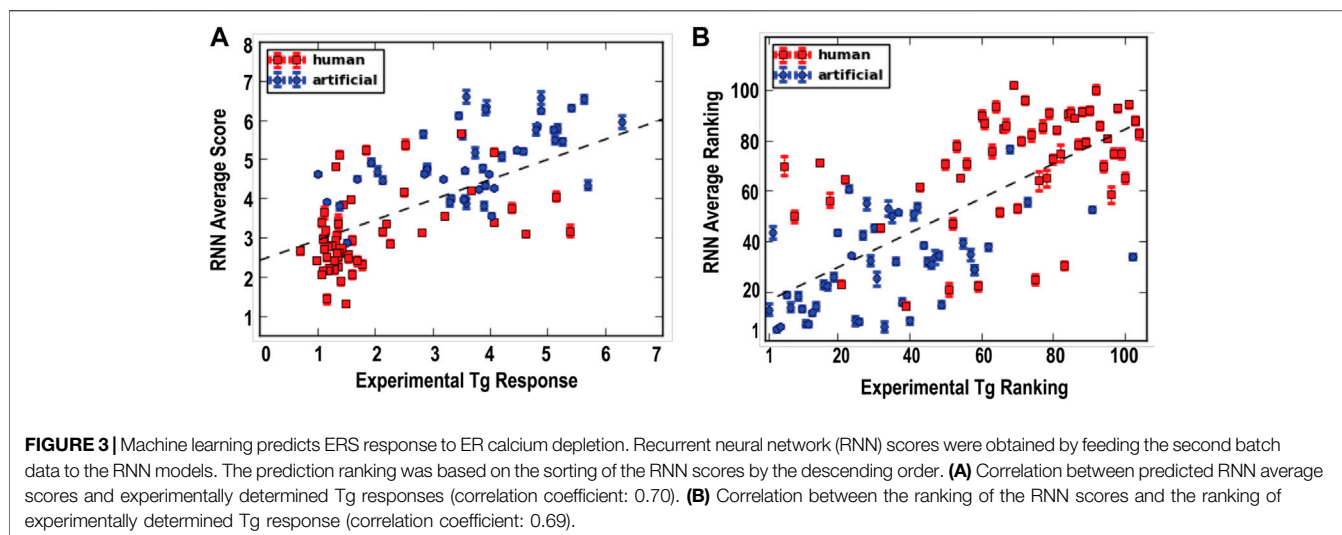
A Machine Learning Algorithm Predicts ERS

Given that the canonical KDEL primary sequence motif cannot reliably predict an ERS, we used a machine learning-based approach to identify potential ERS and their connection to exodosis caused by ER calcium depletion. Using our previously published results of 7-residue tail sequences as the training data (first batch of 95) (Trychta et al., 2018a), we created recurrent neural network (RNN) models with backpropagation and long short-term memory (LSTM) algorithm (Hochreiter and Schmidhuber, 1997). The models were then used to predict 7-residue tail sequences for their responsiveness to Tg treatment (i.e., ER calcium depletion). Based on the predictions by our models, we selected 104 sequences (second batch) for experimental validation. The selected sequences included both those from human proteins and artificial ones that do not exist in the human proteome. In addition to selecting sequences predicted to have strong Tg responses, we also included some sequences predicted to have weak Tg responses to validate our models. The

selected ERS were appended to the C-terminus of GLuc and used for validation in our experimental model of ER calcium depletion.

Comparing the predicted Tg responses (referred to as RNN scores) and the experimentally determined Tg responses, moderate positive correlations were found based on either the response values or their relative rankings, with correlation coefficients of 0.70 and 0.69, respectively (**Figures 3A,B**). Interestingly, a few artificial sequences that were predicted to have high Tg-responses were indeed found to experimentally exhibit higher responses than any of the human sequences tested in the second batch (**Figure 3B**). In addition, our findings further demonstrate that the primary KDEL sequence does not include adequate information to predict ER protein secretion. For example, in both the first and second batches of data, the positively charged Arg and Lys appear to be dominant but the oppositely charged Asp and Glu can be tolerated. This can be observed for KHLEDEL which had an RNN score of 4.1 ± 0.1 and an experimentally determined Tg response of 5.2.

Non-KDEL ERS Form Alternative Interactions with KDEL Receptor

To understand the molecular recognition between the C-terminal ERS and KDEL receptors, we performed molecular modeling studies of selected ERS. Until recently no crystal structures existed for any KDELR, and, as of yet, KDELR1 and KDELR3 remain unresolved. However, a crystal structure of chicken KDEL receptor 2 (KDELR2) demonstrated how the canonical ERS TAEKDEL binds to KDELR2 (Bräuer et al., 2019). In the crystal structure, the deeply buried -1Leu and -2Glu of the bound peptide make extensive interactions with the receptor, and directly contact 10 and 6 residues respectively; in comparison, -3Asp and -4Lys interact with 3 and 4 residues, respectively (**Figure 4A**, **Supplementary Figure S3A**). To investigate these varied extents of interactions in the recognition of the ERS by KDEL receptors, we first built a simulation system of KDELR2 bound with TAEKDEL and embedded in explicit water-lipid bilayer environment, and carried out molecular dynamics (MD) simulations (see

FIGURE 3 | Machine learning predicts ERS response to ER calcium depletion. Recurrent neural network (RNN) scores were obtained by feeding the second batch data to the RNN models. The prediction ranking was based on the sorting of the RNN scores by the descending order. **(A)** Correlation between predicted RNN average scores and experimentally determined Tg responses (correlation coefficient: 0.70). **(B)** Correlation between the ranking of the RNN scores and the ranking of experimentally determined Tg response (correlation coefficient: 0.69).
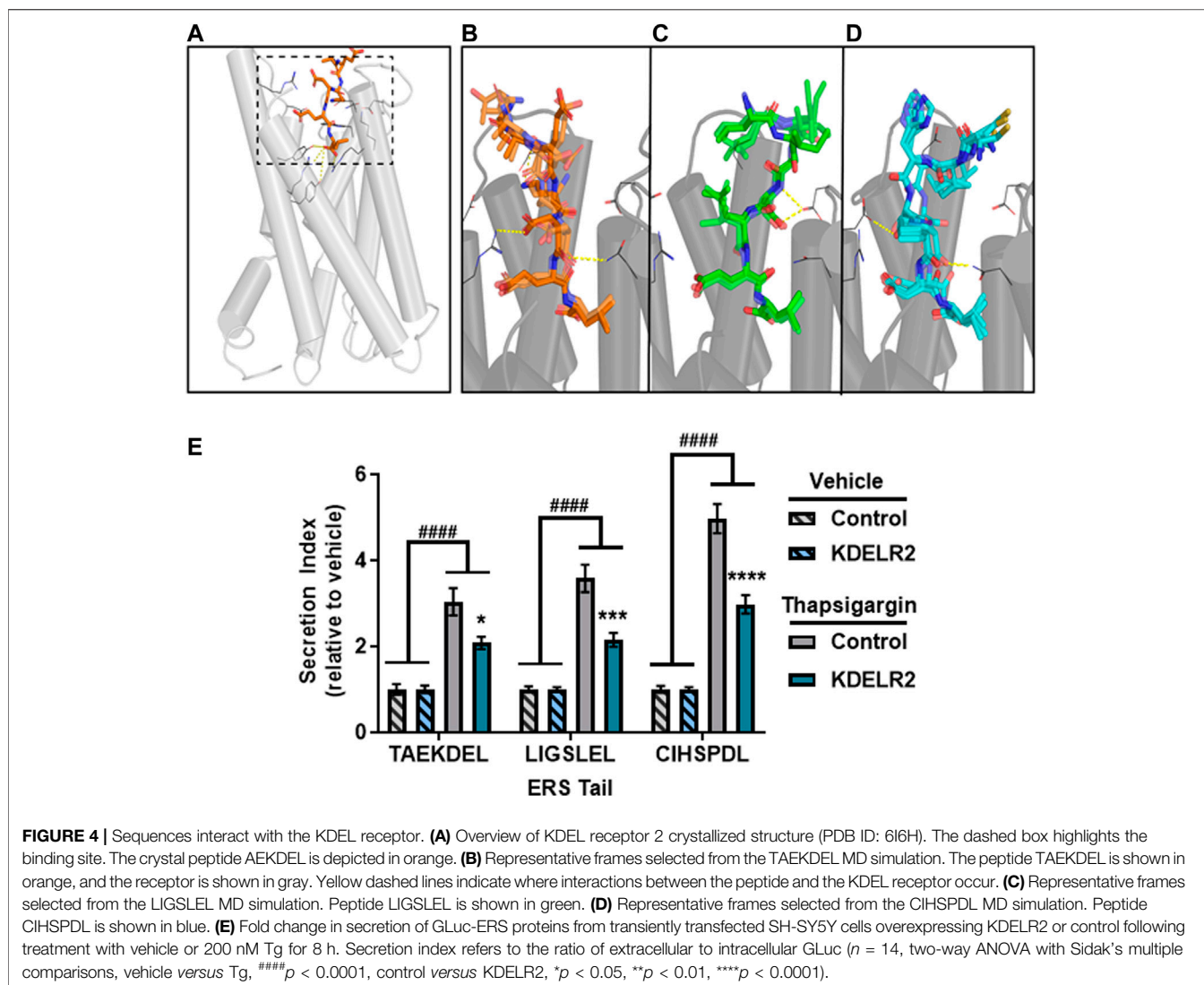
Methods). In our simulations, we observed that the TAEKDEL peptide largely retained the pose found in the crystal structure (**Figure 4B**, **Supplementary Figure S3B**). However, the position of the -3Asp sidechain in the crystal structure appears to be in a high-energy state, which is stabilized only by its interaction with residue Arg169 of KDELR2. In our simulations, while the $\chi 1$ rotamer of -3Asp can be in the gauche + ($\sim +60°$) position for much time as in the structure, it frequently visited the gauche– ($\sim -60°$) position (**Supplementary Figure S4**), a more favored position for Asp (Lovell et al., 2000). In comparison, the $\chi 1$ of -2Glu residue was persistently in its favored gauche- rotamer throughout our simulations (**Supplementary Figure S4**). When $\chi 1$ of -3Asp is in gauche+, it forms an ionic interaction with Arg169; when it is in gauche–, it loses this interaction and becomes dynamic. In comparison, -2Glu persistently forms a salt bridge with Arg169 (**Supplementary Figure S4**).

This finding of the relatively weak interaction between -3Asp and KDELR2 suggests that replacement of Asp at this position can be tolerated for other ERS. We found that alanine substitution at -3Asp had only a modest effect on basal secretion (from data in **Figure 1**). When we calculated the similarity scores of our experimentally verified ERS against KKKKDEL, the most probable amino acid motif derived from our previously identified WebLogo sequence (see also Methods), we found two ERS that are most divergent from KKKKDEL, LIGSLEL and CIHSPDL, do not have Asp at the -3 position. Compared to the RNN score of TAEKDEL, 4.99, those for LIGSLEL are CIHSPDL are 3.26 and 4.56, respectively. To characterize how these two peptides interact with KDELR2, we docked them in KDELR2, similarly built the simulation systems as with KDELR2/TAEKDEL, and carried out MD simulations (see Methods). In our simulations, the -3Leu of LIGSLEL does not make any interaction with the receptor, while -3Pro of CIHSPDL can stably interact with Arg5, Asn60, and Arg169. Even though neither peptide has the -4Lys seen in KKKKDEL, for LIGSLEL, both the sidechain and backbone of -4Ser persistently interact with Asp50 (**Figure 4C; Supplementary Figure S3C**), while the -4Ser of CIHSPDL

interacts with Asp177 keeping the peptide in a stable conformation (**Figure 4D**; **Supplementary Figure S3D**).

The -5, -6, and -7 position residues of the peptides have fewer interactions with the receptor compared to the -1 to -4 residues because they are located at the opening of the binding pocket. However, interestingly for different peptides, residues in the -5, -6, or -7 positions interacted with the different residues of the receptor. Specifically, -7Thr of TAEKDEL interacts with Asp112 of the receptor, while the -6Ile of LIGSLEL and ILE56 of the receptor forms a hydrophobic interaction. For CIHSPDL, -6ILE interacts with four residues, in particular, the hydrophobic interaction with Leu116 of the receptor appears to be a stabilizing force by interacting with -6L (**Supplementary Figure S3**).

Despite their divergence from the canonical KDEL sequence, both peptides can be reasonably adopted in the KDEL receptor binding site, suggesting that KDEL receptors can interact with a variety of ERS tails and that Tg-induced ERS protein secretion is not due to a lack of interaction with KDEL receptors. KDEL receptors normally function to return escaped ER resident proteins from the Golgi to the ER lumen, but following ER calcium depletion KDEL receptors are overwhelmed by the mass efflux of proteins from the ER resulting in increased ERS protein secretion (Trychta et al., 2018a). Overexpression of KDELR2 can attenuate the Tg-induced secretion of GLuc-TAEKDEL, as well as that of GLuc-LIGSLEL and GLuc-CIHSPDL (**Figure 4E**). The increased secretion of LIGSLEL and CIHSPDL constructs following Tg treatment and their predicted interaction with KDELR2 indicate they act in a manner consistent with an ERS secretion phenotype. These data suggest that proteins SFTPB and CNTN2, which contain the predicted ERS LIGSLEL and CIHSPDL, respectively, would be secreted in conditions associated with ER calcium depletion and, in fact, increased secretion of these proteins has been associated with disease (**Supplementary Table S2**).

FIGURE 4 | Sequences interact with the KDEL receptor. **(A)** Overview of KDEL receptor 2 crystallized structure (PDB ID: 6I6H). The dashed box highlights the binding site. The crystal peptide AEKDEL is depicted in orange. **(B)** Representative frames selected from the TAEKDEL MD simulation. The peptide TAEKDEL is shown in orange, and the receptor is shown in gray. Yellow dashed lines indicate where interactions between the peptide and the KDEL receptor occur. **(C)** Representative frames selected from the LIGSLEL MD simulation. Peptide LIGSLEL is shown in green. **(D)** Representative frames selected from the CIHSPDL MD simulation. Peptide CIHSPDL is shown in blue. **(E)** Fold change in secretion of GLuc-ERS proteins from transiently transfected SH-SY5Y cells overexpressing KDELR2 or control following treatment with vehicle or 200 nM Tg for 8 h. Secretion index refers to the ratio of extracellular to intracellular GLuc (n = 14, two-way ANOVA with Sidak's multiple comparisons, vehicle *versus* Tg, [####]$p < 0.0001$, control *versus* KDELR2, [*]$p < 0.05$, [**]$p < 0.01$, [****]$p < 0.0001$).

DISCUSSION

Here we investigate C-terminal ERS sequences and assess how their amino acid sequence can predict protein secretion following ER calcium depletion. We demonstrate that amino acids of the C-terminus beyond the final four residues affect protein secretion in response to ER calcium depletion and form interactions with KDEL receptors. Our recent work suggests that ER calcium depletion causes ERS-containing proteins to be secreted from the cell (Trychta et al., 2018a). The ability to predict whether a protein would be secreted in response to reductions of ER luminal calcium could assist in our understanding of the role of ER luminal proteins in pathological conditions. A loss of essential luminal proteins, i.e., exodosis, could negatively impact ER functioning, while the presence of these normally ER luminal proteins outside of the cell could also change the extracellular environment. A gain of function outside of the cell could be detrimental if the secreted proteins contribute to inflammation or lead to aberrant activation of signaling pathways, but the

extracellular localization of some ERS-containing proteins could have beneficial effects. For example, in rheumatoid arthritis BiP at the cell surface can stimulate tumor necrosis factor (TNF) production to promote inflammation, but BiP can also act as an antigen to stimulate CD8-positive T cells leading to increased production of the anti-inflammatory cytokine IL-10 (Bodman-Smith et al., 2003; Lu et al., 2010). Evaluating whether a protein has a putative ERS may provide mechanistic and etiological understanding of related diseases.

Published data also supports that ERS-containing proteins can saturate KDEL receptors when they leave the ER en masse (Trychta et al., 2018a). In the current molecular modeling study we show that ERS divergent from the canonical KDEL motif can be reasonably accommodated in the KDEL receptor binding site, which is consistent with the observation that overexpression of KDELR2 can attenuate the secretion of proteins with these divergent ERS (**Figure 4**). This evidence of an interaction with KDEL receptors warrants further study to gain a complete picture of the KDEL receptor retrieval pathway and

how ERS differences or KDEL receptor isoforms may influence protein trafficking. Furthermore, the understanding of the presence and action of KDEL receptors on the plasma membrane may provide insight into the mechanism of action of ERS containing proteins that are secreted from the cell following ER calcium depletion (Jia et al., 2020).

The implications of our machine learning algorithm go beyond predicting whether a protein is secreted following ER calcium depletion based on its C-terminal ERS or whether an ERS can dock to KDEL receptors. In this study, we used a Tg-based model of ER calcium depletion, but there are many disease states associated with ER calcium depletion (Mekahli et al., 2011). The ability to predict which proteins may be secreted from the cell in response to calcium depletion can help guide studies of the complex relationship between disruptions in ER homeostasis and diverse pathologies. Our data support that artificial sequences not found in the human proteome could be added to reporters or therapeutic proteins to achieve different degrees of protein secretion in response to calcium depletion. Given our model also predicts proteins that will have an increased extracellular abundance in diseases associated with ER calcium depletion, the development of new biomarkers is feasible. Additionally, for diseases in which ERS proteins have previously been found in increased amounts in extracellular fluids, the role of ER calcium depletion and exodosis in the pathology of the disease may warrant further study. For example, the increased secretion of luciferase reporters with LIGSLEL and CIHSPDL tails suggests their corresponding proteins, SFTPB and CNTN2, respectively, would be elevated in response to ER calcium depletion. Interestingly, the mouse SFTPB, which has the tail CFQTPHL, was also shown to be increased in plasma of a cancer model (Taguchi et al., 2011). Our model predicted CFQTPHL to have an RNN score of 4.36, which is comparable to that of CIHSPDL (4.56). Even though the similarity of CFQTPHL and CIHSPDL is not high, they both have -3Pro, which was found to make extensive interactions with KDELR2 (**Figure 4**), while at the -4 position, Thr is expected to make similar interactions as Ser. In fact, both proteins have been shown to be elevated extracellularly in diseases (**Supplementary Table S2**) and may serve as biomarkers and provide insight into disease mechanism as well as therapeutic approaches to stabilizing ER calcium and preventing exodosis.

We show that ERS divergent from the prototypical sequence "KDEL" are capable of binding to the KDELR and with different interacting residues on the KDELR. The term "KDEL receptor" does not fully reflect the breadth of ligands to which it binds. Our paper helps to further illuminate the diverse ligands that can bind to the KDEL receptors and be retained within the cell but secreted upon decreases in ER calcium. Additionally, the structure used in this study is based off the known structure for KDELR2. The structures for KDELR1 and KDELR3 have not been resolved and may identify additional variations in their ability to bind ERS. Future studies examining the structural differences among the KDELR isoforms and their preferences for different ERS ligands will be necessary to better understand their role in this complex protein retrieval pathway.

Overall, we present a systematic examination of the connection between C-terminal ERS, KDEL receptors, and protein secretion and highlight the combined machine-learning and molecular modeling approaches that can predict and characterize ER resident protein secretion mediated by ER calcium depletion and recognition by the KDEL receptor.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2021.689608/full#supplementary-material

Supplementary Figure 1 | Determining the number of models needed to predict results. **(A)** With 100 times bootstrapping, more models made the correlation estimation between predicted RNN scores and experimental Tg response converge. Each single blue line represents one bootstrapping result. **(B)** The standard error of correlation decreased with an increasing number of models. After the slope reached 0.0003, the calculation was converged with the red line indicating that this convergence occurred with 15 models.

Supplementary Figure 2 | Secretion of GLuc-TAEKDEL fusion proteins from transiently transfected SH-SY5Y cells under vehicle conditions (blue) and after 200 nM Tg for 8 h (purple hash) (mean ± SEM, n = 36 wells, 3 independent experiments). Secretion index refers to the ratio of extracellular to intracellular GLuc and data are not normalized.

Supplementary Figure 3 | The interactions between the bound peptides and KDELR2. **(A)** The blue filled pixels indicate the interactions between the bound peptide AEKDEL residues (left) and the receptor residues (top) revealed in the crystal structure of KDELR2. **(B–D)** The interaction frequencies between the bound peptide TAEKDEL **(B)**, LIGSLEL **(C)**, or CIHSPDL **(D)** residues and KDELR2 residues in our MD simulations. If the shortest heavy-atom distance between two given residues is less than or equal to 4.5 Å, we define them form an interaction. The interaction frequency is defined as the number of MD simulation frames showing the specific interaction divided by the total number of frames in the corresponding MD simulation trajectories. Thus, 0 means no interaction at all, while 1.0 means persistent interactions in the entire simulations.

Supplementary Figure 4 | The -3Asp and -2Glu of TAEKDEL exhibited different interaction patterns to Arg169 during MD simulations of KDELR2. The x-axis is the distances between the -3Asp (blue) or -2Glu (orange) and Arg169 of KDELR2, while the y-axis indicates the $\chi 1$ dihedral angles of -3Asp or -2Glu. Thus, -2Glu can form an ionic interaction to Arg169 of KDELR2 with a heavy-atom distance of ~2.5 Å, while the sidechain of -3Asp is much more dynamic.

REFERENCES

Alanen, H. I., Raykhel, I. B., Luukas, M. J., Salo, K. E. H., and Ruddock, L. W. (2011). Beyond KDEL: the Role of Positions 5 and 6 in Determining ER Localization. *J. Mol. Biol.* 409, 291–297. doi:10.1016/j.jmb.2011.03.070

Bodman-Smith, M. D., Corrigall, V. M., Kemeny, D. M., and Panayi, G. S. (2003). BiP, a Putative Autoantigen in Rheumatoid Arthritis, Stimulates IL-10-producing CD8-Positive T Cells from normal Individuals. *Rheumatology (Oxford)* 42, 637–644. doi:10.1093/rheumatology/keg204

Booth, C., and Koch, G. L. E. (1989). Perturbation of Cellular Calcium Induces Secretion of Luminal ER Proteins. *Cell* 59, 729–737. doi:10.1016/0092-8674(89)90019-6

Bräuer, P., Parker, J. L., Gerondopoulos, A., Zimmermann, I., Seeger, M. A., Barr, F. A., et al. (2019). Structural Basis for pH-dependent Retrieval of ER Proteins from the Golgi by the KDEL Receptor. *Science* 363, 1103–1107. doi:10.1126/science.aaw2859

Collins, J. E., Wright, C. L., Edwards, C. A., Davis, M. P., Grinham, J. A., Cole, C. G., et al. (2004). A Genome Annotation-Driven Approach to Cloning the Human ORFeome. *Genome Biol.* 5, R84. doi:10.1186/gb-2004-5-10-r84

Fliegel, L., Newton, E., Burns, K., and Michalak, M. (1990). Molecular Cloning of cDNA Encoding a 55-kDa Multifunctional Thyroid Hormone Binding Protein of Skeletal Muscle Sarcoplasmic Reticulum. *J. Biol. Chem.* 265, 15496–15502. doi:10.1016/s0021-9258(18)55423-5

Glembotski, C. C., Thuerauf, D. J., Huang, C., Vekich, J. A., Gottlieb, R. A., and Doroudgar, S. (2012). Mesencephalic Astrocyte-Derived Neurotrophic Factor Protects the Heart from Ischemic Damage and Is Selectively Secreted upon Sarco/endoplasmic Reticulum Calcium Depletion. *J. Biol. Chem.* 287, 25893–25904. doi:10.1074/jbc.m112.356345

Henderson, M. J., Richie, C. T., Airavaara, M., Wang, Y., and Harvey, B. K. (2013). Mesencephalic Astrocyte-Derived Neurotrophic Factor (MANF) Secretion and Cell Surface Binding Are Modulated by KDEL Receptors. *J. Biol. Chem.* 288, 4209–4225. doi:10.1074/jbc.m112.400648

Henderson, M. J., Wires, E. S., Trychta, K. A., Richie, C. T., and Harvey, B. K. (2014). SERCaMP: a Carboxy-Terminal Protein Modification that Enables Monitoring of ER Calcium Homeostasis. *MBoC* 25, 2828–2839. doi:10.1091/mbc.e14-06-1141

Henderson, M. J., Wires, E. S., Trychta, K. A., Yan, X., and Harvey, B. K. (2015). Monitoring Endoplasmic Reticulum Calcium Homeostasis Using a Gaussia Luciferase SERCaMP. *J. Visualized Exp.* (103), e53199. doi:10.3791/53199

Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.* 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735

Hsu, V. W., Shah, N., and Klausner, R. D. (1992). A Brefeldin A-like Phenotype Is Induced by the Overexpression of a Human ERD-2-like Protein, ELP-1. *Cell* 69, 625–635. doi:10.1016/0092-8674(92)90226-3

Jia, J., Yue, X., Zhu, L., Jing, S., Wang, Y., Gim, B., et al. (2020). KDEL Receptor Is a Cell Surface Receptor that Cycles between the Plasma Membrane and the Golgi *via* Clathrin-Mediated Transport Carriers. *Cell Mol. Life Sci.* 78, 1085–1100. doi:10.1007/s00018-020-03570-3

Lewis, M. J., and Pelham, H. R. B. (1990). A Human Homologue of the Yeast HDEL Receptor. *Nature* 348, 162–163. doi:10.1038/348162a0

Lewis, M. J., and Pelham, H. R. B. (1992a). Ligand-induced Redistribution of a Human KDEL Receptor from the Golgi Complex to the Endoplasmic Reticulum. *Cell* 68, 353–364. doi:10.1016/0092-8674(92)90476-s

Lewis, M. J., and Pelham, H. R. B. (1992b). Sequence of a Second Human KDEL Receptor. *J. Mol. Biol.* 226, 913–916. doi:10.1016/0022-2836(92)91039-r

Lovell, S. C., Word, J. M., Richardson, J. S., and Richardson, D. C. (2000). The Penultimate Rotamer Library. *Proteins* 40, 389–408. doi:10.1002/1097-0134(20000815)40:3<389:aid-prot50>3.0.co;2-2

Lu, M.-C., Lai, N.-S., Yu, H.-C., Huang, H.-B., Hsieh, S.-C., and Yu, C.-L. (2010). Anti-citrullinated Protein Antibodies Bind Surface-Expressed Citrullinated Grp78 on Monocyte/macrophages and Stimulate Tumor Necrosis Factor α Production. *Arthritis Rheum.* 62, 1213–1223. doi:10.1002/art.27386

Mazzarella, R. A., Marcus, N., Haugejorden, S. M., Balcarek, J. M., Baldassare, J. J., Roy, B., et al. (1994). ERp61 Is GRP58, a Stress-Inducible Luminal Endoplasmic Reticulum Protein, but Is Devoid of Phosphatidylinositide-Specific Phospholipase C Activity. *Arch. Biochem. Biophys.* 308, 454–460. doi:10.1006/abbi.1994.1064

Mazzarella, R. A., Srinivasan, M., Haugejorden, S. M., and Green, M. (1990). ERp72, an Abundant Luminal Endoplasmic Reticulum Protein, Contains Three Copies of the Active Site Sequences of Protein Disulfide Isomerase. *J. Biol. Chem.* 265, 1094–1101. doi:10.1016/s0021-9258(19)40163-4

Mekahli, D., Bultynck, G., Parys, J. B., de Smedt, H., and Missiaen, L. (2011). Endoplasmic-reticulum Calcium Depletion and Disease. *Cold Spring Harb Perspect. Biol.* 3, a004317. doi:10.1101/cshperspect.a004317

Munro, S., and Pelham, H. R. B. (1987). A C-Terminal Signal Prevents Secretion of Luminal ER Proteins. *Cell* 48, 899–907. doi:10.1016/0092-8674(87)90086-9

Orci, L., Stamnes, M., Ravazzola, M., Amherdt, M., Perrelet, A., Söllner, T. H., et al. (1997). Bidirectional Transport by Distinct Populations of COPI-Coated Vesicles. *Cell* 90, 335–349. doi:10.1016/s0092-8674(00)80341-4

Raykhel, I., Alanen, H., Salo, K., Jurvansuu, J., Nguyen, V. D., Latva-Ranta, M., et al. (2007). A Molecular Specificity Code for the Three Mammalian KDEL Receptors. *J. Cell Biol.* 179, 1193–1204. doi:10.1083/jcb.200705180

Roos, K., Wu, C., Damm, W., Reboul, M., Stevenson, J. M., Lu, C., et al. (2019). OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. *J. Chem. Theor. Comput.* 15, 1863–1874. doi:10.1021/acs.jctc.8b01026

Sherman, W., Day, T., Jacobson, M. P., Friesner, R. A., and Farid, R. (2006). Novel Procedure for Modeling Ligand/receptor Induced Fit Effects. *J. Med. Chem.* 49, 534–553. doi:10.1021/jm050540c

Taguchi, A., Politi, K., Pitteri, S. J., Lockwood, W. W., Faça, V. M., Kelly-Spratt, K., et al. (2011). Lung Cancer Signatures in Plasma Based on Proteome Profiling of Mouse Tumor Models. *Cancer cell* 20, 289–299. doi:10.1016/j.ccr.2011.08.007

Thastrup, O., Cullen, P. J., Drobak, B. K., Hanley, M. R., and Dawson, A. P. (1990). Thapsigargin, a Tumor Promoter, Discharges Intracellular Ca2+ Stores by Specific Inhibition of the Endoplasmic Reticulum Ca2(+)-ATPase. *Proc. Natl. Acad. Sci. U S A.* 87, 2466–2470. doi:10.1073/pnas.87.7.2466

Trychta, K. A., Bäck, S., Henderson, M. J., and Harvey, B. K. (2018a). KDEL Receptors Are Differentially Regulated to Maintain the ER Proteome under Calcium Deficiency. *Cell Rep.* 25, 1829–1840.e6. doi:10.1016/j.celrep.2018.10.055

Trychta, K. A., Heathward, E. J., Sulima, A., Bäck, S., Farokhnia, M., Richie, C. T., et al. (2018b). Extracellular Esterase Activity as an Indicator of Endoplasmic Reticulum Calcium Depletion. *Biomarkers* 23, 756–765. doi:10.1080/1354750x.2018.1490968

Wu, Z., Newstead, S., and Biggin, P. C. (2020). The KDEL Trafficking Receptor Exploits pH to Tune the Strength of an Unusual Short Hydrogen Bond. *Sci. Rep.* 10, 16903. doi:10.1038/s41598-020-73906-3