Massively parallel interrogation of protein fragment secretability using SECRiFY reveals features influencing secretory system transit

Boone, M.[1,2,7,*]; Ramasamy, P.[1,3,4,5,6], Zuallaert, J.[1,2,8,9], Bouwmeester, R.[1,3], Van Moer, B.[1,2], Maddelein, D.[1,3], Turan, D.[1,3], Hulstaert, N.[1,3], Eeckhaut, H.[1,2], Vandermarliere, E.[1,3], Martens, L.[1,3], Degroeve, S.[1,3], De Neve, W.[8,9], Vranken, W.[4,5,6], Callewaert, N.[1,2,*]

1. Center for Medical Biotechnology, VIB, Zwijnaarde, Belgium

2. Department of Biochemistry and Microbiology, Faculty of Sciences, Ghent University, Ghent, Belgium

3. Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium

4. Structural Biology Brussels, VUB, Brussels, Belgium

5. Structural Biology Research Center, VIB, Brussels, Belgium

6. Interuniversity Institute of Bioinformatics in Brussels (IB)2, ULB-VUB, Brussels, Belgium

7. Present address: Department of Biochemistry and Biophysics, UCSF, San Francisco, CA, USA

8. Center for Biotech Data Science, Ghent University Global Campus, Songdo, Incheon, South Korea

9. IDLab, ELIS, UGent, Ghent, Belgium

* corresponding authors

Abstract

While transcriptome- and proteome-wide technologies to assess processes in protein biogenesis are now widely available, we still lack global approaches to assay post-ribosomal biogenesis events, in particular those occurring in the eukaryotic secretory system. We here developed a method, SECRiFY, to simultaneously assess the secretability of >10^5 protein fragments by two yeast species, *S. cerevisiae* and *P. pastoris,* using custom fragment libraries, surface display and a sequencing-based readout. Screening human proteome fragments with a median size of 50 - 100 amino acids, we generated datasets that enable datamining into protein features underlying secretability, revealing a striking role for intrinsic disorder and chain flexibility. SECRiFY is the first methodology that generates sufficient amounts of annotated data for advanced machine learning methods to deduce secretability predictors. The finding that secretability is indeed a learnable feature of protein sequences is of significant impact in the broad area of recombinant protein expression and *de novo* protein design.

<u>Main text</u>

Introduction

The eukaryotic secretory system processes roughly a quarter of the proteome[1–3], ensuring correct folding, assembly, and delivery of proteins to the extracellular environment, the plasma membrane, or membrane-bound organelles[4–6]. Model secretory cargos such as yeast carboxypeptidase Y (CPY), α-1 antitrypsin (AAT), transthyretin (TTR), the Cystic Fibrosis transmembrane conductance regulator (CFTR), and the vesicular stomatitis virus G protein (VSVG) have been instrumental in understanding the function and regulation of many ER- or Golgi-resident proteins (for instance [7–11]); yet, the precise features that enable or prevent secretory system transit of the thousands of other secretory proteins remain obscure. For example, it is generally unknown which chaperones are critical for assisting the folding of the different types of secretory protein domains, what sequence or structural features control ER export kinetics, or what determines glycan modification by Golgi glycosyltransferases. Studies that examine a broad range of proteins passing through the secretory system are integral to understanding how multiple processes integrate to produce the full set of secretory proteins. Unfortunately, most current approaches are unsuited to study a comprehensive range of protein folds after entry in the ER. Mass spectrometry (MS)-based proteomics is the predominant approach for the interrogation of post-translational events, but despite many technological advances, its breadth and depth is limited and decreases steeply with sample complexity; in routine MS setups, generally less than 70% of all transcribed mammalian protein-coding genes are detected[12,13], and full protein coverage is rarely achieved.

In practice, the absence of such an integrated picture of the secretory system is most apparent in the unpredictability of heterologous protein secretion. Indeed, four decades after the recombinant DNA revolution, obtaining detectable levels of functional protein in a particular heterologous host system, secretory or not, is still principally a process of trial and error. Statistics from worldwide Structural Genomics Consortia initiatives suggest that the likelihood of progression from cloned target to soluble, purifiable protein lies between 20-50%, even with target selection[14,15]. These low and unpredictable expression success rates slow down progress in the many fields of basic and applied life science where recombinant protein production comes into play. Although expression screening of parallel constructs or variant libraries of a protein of interest to increase heterologous protein expression success rates has gained momentum[16–19], it has to be repeated for each new target. Additionally, alternative comprehensive strategies to assess heterologous expression across entire proteomes have generally been limited to intracellular expression in *E. coli*, small proteomes, and cumbersome clone-by-clone strategies[20–23]. As the availability of large-scale protein expressability data will help to demystify which and why proteins fail to express, new methods for measuring expression in high throughput are needed.

We here developed an approach to evaluate the secretory potential ('secretability') of proteins on a proteome-wide scale. SECRiFY (SECretability screening of Recombinant Fragments in Yeast) combines yeast surface display screening of protein libraries and a deep sequencing readout, enabling the systematic identification of heterologous polypeptides that can pass (or evade) the secretory quality control checkpoints of the yeast ER, Golgi, secretory vesicles and plasma membrane, and be secreted. As a first fundamental question to be addressed using this new methodology, we asked whether, given a particular sequence of a protein fragment, we could 1) learn what features contribute to its secretability and 2) generate machine-learned classifiers that predicts its secretability. Hence, we fragmented the human proteome and screened these fragments for secretability in two different yeast species, *Saccharomyces cerevisiae* and *Pichia pastoris* (*Komagataella phaffii*). We generated a large repository of more than 20,000 experimentally determined yeast-producible human protein fragments, available to the research community at https://iomics.ugent.be/secrify. These datasets of experimentally verified secretable fragments will assist experimental design of recombinant protein secretion. We used them to train (deep) machine learning models for secretability prediction, which unveiled sequence and structural determinants of productive secretory system transit, highlighting the utility of SECRiFY to provide further insight into the basic mechanisms of secretory processing. More application-focused implementation of SECRiFY (focusing on fixed-boundary protein domains or multi-domain fragments) should enable to generate databases of experimentally validated secretable native protein domain fragments, which could tremendously speed up experimental protein expression in many fields of study.

Results

Normalized fragment libraries for screening at domain-level resolution

Multi-domain proteins often fail to express or secrete in their entirety due to local issues with misfolding of particular protein areas, translation inhibitory sequences, protease susceptibility, the absence of stabilizing interaction partners or modifications, or toxicity. The structural, functional and evolutionary modularity of proteins in domains, however, implies that individual expression of certain protein parts, especially domains, can often nonetheless be achieved. Chopping up difficult proteins into experimentally tractable fragments has been exploited by structural biologists for years, both in rational target design as well as in random library screens for soluble expression[24–28]. Moreover, screening of protein domains or fragments can provide valuable information that is not immediately attainable or obvious from screens with full-length proteins[29]. Some domain-focused interactome studies, for example, have allowed immediate

103 delineation of the minimal interacting regions and the detection of more interactors without
104 increasing the number of false positives[30]. We thus rationalized that screening libraries of
105 domains or domain-sized polypeptides, rather than full-length proteins, would allow for a higher
106 resolution measurement of secretability across proteomes, and facilitate the identification of
107 sequence or structural determinants of secretion.

108 Domain boundary prediction, however, is notoriously inaccurate, and even with a reliable
109 estimate, small variations in the exact N- and C-terminus of the fragment can lead to dramatic
110 differences in expressability[31]. Random approaches, on the other hand, can generate libraries
111 of fragments that encompass most domains of a proteome by sheer oversampling. We
112 therefore designed and built directional, randomly fragmented cDNA libraries covering the
113 human transcriptome with fragments coding for approx. 50-100 amino acids, which is the
114 median domain size of human proteins (**Fig. 1a, c**). Due to the large dynamic range of mRNA
115 transcripts in human cells (in abundance of 4 orders of magnitude), however, capturing the full
116 diversity of fragments would require unfeasibly large libraries, even at 100 bp resolution (**Fig.
117 1b**). We therefore reduced fragment abundance differences by relying on the second-order
118 kinetics of nucleic acid rehybridization after denaturation, and subsequent digestion with the
119 Kamchatka crab duplex-specific nuclease (DSN)[32,33] (**Fig. 1c-d**). More abundant DNA species
120 rehybridize faster and are therefore digested first; as such, even a single round of normalization
121 substantially reduces abundance differences between DNA fragments (**Fig. 1g**). Crucially, this
122 allowed us to downsize the libraries to a scale that is feasibly compatible with downstream
123 cDNA library cloning and yeast transformation efficiencies (+/- $5*10^6$ - $5*10^7$).

124 To ensure directionality, random primers were tagged with a rare-cutter restriction site (PacI),
125 which is distinct from the SfiI site incorporated in the library adapters (**Fig. 1c**). We initially
126 observed that the random primer tag sequence is susceptible to degradation due to endo- and
127 exonuclease activity of the *E. coli* DNA polymerase I during second strand synthesis[34–37]. As a
128 result, less than 20% of fragment sequences contained a full-length PacI site, negatively
129 affecting ligation into the surface display vector (**Fig. 1e**). Both nuclease-resistant
130 phosphorothioate bonds and buffer sequences could partially protect the tag from degradation,
131 and for the final library design, we settled on the primer where protection efficiency was
132 maximal (**Fig. 1e, grey bar**).

133 Tag composition also affected abundance normalization efficiencies. In an earlier design with
134 a GC-rich tag, normalization was less effective (**Fig. 1f**) than the design with PacI tag (**Fig.
135 1g**), where an approx. 1000-fold normalization could routinely be obtained. The tag sequence
136 is present on all sequence fragments, and most likely, when using a GC-rich tag,

137 rehybridization kinetics (and therefore degradation) is dominated by the tag rather than by the
138 sequence of the fragments themselves.

139 In all, this library construction protocol allows for efficient capture of protein-coding fragments
140 tiled along eukaryotic transcriptomes. It is, to our knowledge, the first time that an effective
141 method for normalization of tagged random-primed cDNA fragment libraries is reported, and it
142 should find many applications in areas where the protein-coding potential of a cell needs to be
143 effectively covered in expression libraries.

144

145 *SECRiFY as a platform for secretability screening in yeast*

146 Relying on the sophisticated quality control (QC) machinery of the eukaryotic secretory system,
147 which ensures efficient degradation of unstable or misfolded proteins before reaching the
148 plasma membrane[19,20], we further reasoned that surface display could be used as a proxy for
149 productive secretion. As such, once cloned into the surface display vector and transferred to
150 yeast, library polypeptides are directed to the secretory system by an N-terminal secretory
151 leader sequence derived from the yeast α mating factor (MFα1 prepro), and furthermore on
152 the yeast cell wall via C-terminal fusion to the GPI-anchoring region of the *S. cerevisiae* cell
153 wall protein Sag1 (**Fig. 2a-b**). Fragments for which the fragment-Sag1 fusion successfully
154 passes (or escapes) secretory system QC without proteolytic degradation are recognized
155 through their N- and C-terminal epitope tags (FLAG and V5, resp.), and are segregated from
156 the rest using iterations of high-efficiency magnetic- and fluorescence-activated cell sorting
157 (MACS/FACS) (**Fig. 2b-c**). Finally, fragment identification and classification are achieved by
158 deep sequencing of fragment amplicons from the unsorted and sorted cell populations. In
159 short, SECRiFY assesses secretability, i.e. the potential of a polypeptide to transit through the
160 secretory system of ER, Golgi, vesicles, and PM without degradation, in a manner that is
161 independent of the original endogenous localization of the protein of interest. For the present
162 study, we focus on basic principles of secretability. While in practice, any protein-coding mRNA
163 pool is compatible with SECRiFY, considering its biomedical importance and structural
164 complexity, we here focused on the human proteome for our screens, as encoded by the
165 transcriptome of various human cell lines.

166 We first benchmarked the method by building a $1.96*10^6$ clone fragment library of the
167 HEK293T transcriptome and performed triplicate screens in *S. cerevisiae*. On average, 1.76%
168 $\pm$ 0.12% of library cells displayed a fragment with an intact N-terminus (FLAG-tag) and intact
169 C-terminus (V5-tag) (**Supplementary Fig. 1**). Accounting for a 1/9 chance of up- and
170 downstream in-frame cloning, this means that approximately 15.8% of in-frame fragments are

171 detectably displayed and hence, potentially secretable. After a 32-fold enrichment of these
172 double positive cells through a single round of MACS and two subsequent rounds of FACS
173 (**Supplementary Fig. 1**), both pre- and post-sort population were sequenced at a per-base
174 average coverage of minimally 150 reads. On average, $1.12*10^6$ unique fragments/replicate
175 were detected, covering on average 26.45% ± 0.86% of the human canonical transcriptome
176 with at least three reads (**Supp. Table 1-3**). To assess the secretion-predictive value of the
177 method, we picked random clones from the sorted population of a single experiment
178 (**Supplementary Fig. 2, Supplementary Table 4**) and tested the secretion of their encoded
179 fragments when not fused to the anchor protein Sag1. The N- or C-terminal tags of 18/20 (90%)
180 fragments could be reliably detected on western blot from the growth medium, and for 16/20
181 (80%) fragments, both tags were recognized (**Fig. 1d, Supp. Table 5**). As such, fragments
182 displayed by sorted cells are indeed 'secretable' with a high probability. We further classified
183 fragments into those that were enriched (also referred to as secretable) and those that were
184 passively depleted (hence, not detected as secretable) by sorting, setting a cut-off on the
185 enrichment factor ($E\ factor = log2\frac{FPTM_{sorted}}{FPTM_{unsorted}}$) at 1 and -1, respectively, reflecting a minimal
186 2-fold increase and decrease in normalized sequence read counts after sorting. Of 170,226 in-
187 frame fragments commonly detected in the three experiments, 6.83% were consistently
188 enriched in all three replicates, and 80.21% consistently depleted (**Supp. Table 6,**
189 **Supplementary Fig. 3**). Thus, using this metric, these screens were reproducible with an
190 87.03% concordance between replicates. These final stratified groups of fragments, which
191 were concordantly enriched or depleted, will further be referred to as secretable and depleted,
192 respectively.

193 Since we only performed positive selection for secretable fragments during screening, the
194 depleted fraction contains only passively depleted fragments, and the negative predictive value
195 is relatively low (40-73%, **Supplementary Fig. 4**). In light of this, interpretation of features
196 affecting secretion must focus on those that affect secretability, and not non-secretability.
197 However, as there are ± 15 times as many fragments (data points) in this depleted set, this
198 relatively low negative predictive value still provides for sufficient signal to allow machine
199 learning methods to learn (see below).

200 Although we initially tested our method in the model yeast *S. cerevisiae*, in practice, the
201 methylotroph *Pichia pastoris* (*Komagataella phaffii*) is an increasingly popular choice of host
202 for recombinant protein production. Mostly, this has been attributed to this yeast's formidable
203 capacity for high-density growth, the secretion of relatively few endogenous secreted proteins,
204 and the availability of very tightly repressed and extraordinarily strong inducible promoters
205 derived from the yeast methanol metabolism genes[38,39]. Key to the adaptation of SECRiFY for

206 use in *P. pastoris* was the development of a modified protocol for high-efficiency large-scale
207 *P. pastoris* transformation, which resulted in an improvement in transformation efficiency of 2
208 - 3 orders of magnitude (**Online Methods** and **Supplementary Fig. 5**). While we previously
209 observed a slight bias towards detecting small fragments in both enriched and depleted
210 classes in our *S. cerevisiae* pilot screen, reducing the number of PCR amplification cycles
211 during library generation for sequencing largely eliminated this trend, although small skews
212 occurring during both cloning and sequencing were still observed (**Supplementary Fig. 6**). For
213 the *P. pastoris* screens presented here, we first generated a new fragment library with slightly
214 larger fragment insert sizes from the pooled transcriptome of four different human cell lines
215 (SK-N-SH_RA, GM12878, HepG2 and MCF-7) originating from diverse human tissues (brain,
216 blood, liver and breast) and selected to maximize the number of expressed human genes
217 based on ENCODE transcriptome data[40]. Our high-efficiency transformation to *P. pastoris*
218 generated a library with an estimated diversity of $9.8*10^6$ clones. Averaged over three replicate
219 screens, 4.06% ± 0.68% of cells from this library were $FLAG^+V5^+$ (**Supplementary Fig. 7**),
220 which, accounting for the frequent presence of multi-copy inserts (**Supplementary Fig. 8**),
221 suggests that 12.18% of in-frame fragments are displayed and hence, potentially secretable.
222 Sequencing the fragments of unsorted cells and cells sorted after 1 round of MACS and 1
223 round of FACS, we detected ± 1.5 million unique fragments per replicate, either in the enriched
224 protein-displaying library, in the non-enriched starting library, or in both, covering 38.38% ±
225 2.25% of the human canonical transcriptome with at least three reads (**Suppl. Table 7-9**). Of
226 the 215,004 in-frame fragments detected in all three of the replicates, 4.84% were classified
227 as consistently enriched, and 65.75% as consistently depleted, leading to a 71% concordance
228 between replicates (**Suppl. Table 10, Supplementary Fig. 9**).

229 Overall, these data show that SECRiFY is a reproducible and a reliable method to estimate
230 the secretability of protein fragments. This dataset now represents by far the largest resource
231 on eukaryotic secretability of protein fragments. To enable researchers to easily access and
232 analyze the data from the screens in this study, we built an interactive database
233 (http://iomics.ugent.be/secrify/) allowing visualization of the protein fragments detected in
234 these screens and their mapping on available PDB structures (**Supplementary Fig. 10,**
235 Figshare links to data in **Online Methods**).

236

237 *Secretable fragments are more flexible and disordered*

238 Just as cytosolic protein expression is influenced by a variety of DNA, mRNA, and protein
239 sequence or structural features and their complex interplays[41–44], secretion of polypeptides will

240 depend on a combination of multiple parameters, some of which are related to the unique
241 environment and QC machinery of the ER and beyond. Even already at the simple level of
242 general averaged parameters over our secretable vs depleted protein fragment collections,
243 several intriguing observations emerged from our data.

244 We first examined whether secretable fragments differed from depleted ones in their probability
245 to form secondary structures. To maximize the accuracy of feature prediction, fragments were
246 filtered for size and exact match to Uniprot proteins, and condensed to an unambiguous subset
247 of consolidated sequences in order to reduce sequence redundancy (see **Online methods**).
248 Secondary structure prediction of this consolidated subset shows that secretable fragments
249 most prominently have a lower propensity to form α-helical structures ($p= 2.95*10^{-124}$, Mann-
250 Whitney-Wilcoxon test) (**Fig. 3a, Supplementary Fig. 11a)**. Indeed, when clustering
251 overlapping sequences to representative fragments and mapping these to solved structures in
252 PDB (roughly 50% of representative fragments, **Supplementary Fig. 12**), secretability
253 similarly inversely correlates with α-helical content (**Fig. 3b, Supplementary Fig. 13a**) ($p=$
254 $2.35*10^{-8}$, Mann-Whitney-Wilcoxon test). In contrast, differences in β-sheet content are only
255 minimal (**Fig. 3a-b**).

256 Since secretable fragments also tend to more readily form random coils than depleted
257 fragments, based on secondary structure predictions ($p= 1.99*10^{-197}$, Mann-Whitney-Wilcoxon
258 test) as well as PDB mapping ($p= 1.24*10^{-4}$, Mann-Whitney-Wilcoxon test) (**Fig. 3a-b**), we
259 further examined how backbone dynamics and intrinsic disorder relate to secretability. As
260 predicted using Dynamine[45,46], secretable fragments are distinctly more flexible ($p= 5.08*10^{-}$
261 ^{118}, Mann-Whitney-Wilcoxon test) (**Fig. 3c**, **Supplementary Fig. 14a**). Disorder calculations on
262 the full secretable vs depleted sets with RAPID[47] also confirmed a higher average intrinsic
263 disorder content in secretable fragments ($p < 2.2*10-^{16}$, Mann-Whitney-Wilcoxon test) **(Fig. 3d**,
264 **Supplementary Fig. 15**). In line with this, on average, fragments from both subsets appear
265 compositionally biased. A larger fraction of secretable fragments has a higher proportion of
266 negatively charged residues and prolines, and a tendency towards lower hydrophobicity
267 (**Supplementary Fig. 16**). Possibly, this increase disorder in secretable fragments reflects how
268 unstructured fragment sequences that lack typically exposed hydrophobic amino acids are
269 missed by ER chaperones and can subsequently travel downstream. This is particularly
270 striking since endogenous secretory system proteins in both human and yeast are, on average,
271 less disordered than the whole proteome, both when considering overall disorder content ($p <$
272 $2.2*10^{-16}$ and $p= 2.46*10^{-5}$ resp., Fisher exact test) as well as absolute number of disordered
273 amino acids ($p < 2.2*10^{-16}$ and $p= 9.44*10^{-10}$ resp., Fisher exact test) (**Suppl. Table 11-12**),
274 suggesting evolutionary counterselection.

275 Increasing the fidelity of the above findings, all feature enrichment observations were
276 reproduced in the *P. pastoris* SECRiFY screens (**Supplementary Tables 7-10,**
277 **Supplementary Figs. 6-18**). In addition, our conclusions remained unchanged when choosing
278 alternative criteria for defining secretable vs depleted fragment sets, illustrating the robustness
279 of our observations.

280

281 *In silico secretability prediction with machine learning*

282 For the first time, our SECRiFY method generates secretability data at a scale at which training
283 of predictive machine learning classifiers becomes feasible. To study the presence of
284 discriminatory features in the dataset, we explored two distinct approaches: one based on
285 feature engineering together with gradient boosted decision tree modeling[48], and a deep
286 learning approach based on convolutional neural networks (CNNs)[49].

287 The gradient boosting classifier requires a fixed input size. Therefore, a series of manually
288 engineered input features were proposed, based on physicochemical properties, sequence
289 length and amino acid frequencies. Ten individual classifiers were trained using different
290 properties, and an ensemble of those was constructed using another gradient boosted
291 classifier taking the outputs of the individual classifiers as input. The deep learning approach
292 involved a CNN taking a one-hot encoding as input, followed by three blocks of convolutional,
293 max pooling and dropout layers. We explored different strategies to deal with the variable input
294 size, as this is not supported by standard CNN architectures. A global max pooling layer
295 yielded the best overall results. This layer is finally connected to a dense layer, followed by an
296 output layer with a softmax.

297 Fragments shorter than 50 amino acids were removed from both the *S. cerevisiae* and *P.*
298 *pastoris* datasets, as those are likely not long enough to properly fold, which mitigates their
299 relevance. Using a restrictive 10-fold cross-validation scheme, where we made sure that
300 protein fragments originating from the same gene were included in the same fold, we compared
301 the classifiers based on the area under the receiver operating characteristic curve
302 (AUROC). Gradient boosting achieved an AUROC of 0.781 and 0.772 on the *S. cerevisiae*
303 and *P. pastoris* datasets, respectively, whereas the CNNs achieved AUROCs of the same
304 magnitude, 0.779 and 0.768 (**Fig. 4a**). Classification results of both classifiers thus confirmed
305 the presence of distinctive features within both secretable and depleted subsets of the data.
306 We observed a strong correlation between the predicted values for the two approaches, with
307 Pearson correlation coefficients of 0.810 and 0.887 on the respective datasets (**Fig. 4b**), which
308 suggests that the two models learned to use similar distinctive features in the data.

309 Feature importance analysis using attribution methods led to compelling insights in the
310 decisions of the CNN (**Fig. 4c-d**). Aggregation of individual attribution maps on the amino acid
311 residue level indicated that the influence of individual residues on secretability is largely
312 independent of their position in the sequence. Strikingly, there is a positive bias towards
313 smaller residues, in line with our biophysical predictions that random coils are more readily
314 formed in secretable fragments. Negatively charged residues also seemed to substantially
315 contribute to secretability, confirming the pattern we picked up looking at simple averaged
316 parameters across the full dataset. Similarly, a negative bias was observed towards all
317 hydrophobic amino acids, affirming our earlier observations.

318

319 *Secretable fragments are enriched in certain folds and domains*

320 Other features that affect folding in the ER might influence secretability. Although increased
321 presence of N-glycosylation sequons or uneven number of cysteines could increase ER
322 retention of fragments, we did not observe clear differences in number of Cys or N-
323 glycosylation sequons (**Supplementary Fig. 16f-g**). Furthermore, secreted and depleted
324 fragments do not substantially differ in their predicted propensity to collapse into folded
325 structure (EFoldMine[50] prediction, **Supplementary Fig. 14b**). Nonetheless, in select cases
326 where depleted and enriched fragments overlap in sequence on the same protein, the
327 presence or absence of regions that are most likely to fold rapidly often correlated with
328 secretability (**Fig. 3e**).

329 Beyond these parameters, we noticed clear differences in protein folds and domain
330 architectures represented in both fragment groups. Of those fragments that mapped to known
331 structures, secretable fragments are enriched in distorted sandwich and β complex folds
332 compared to depleted fragments, suggesting that these folds are potentially more stable in the
333 secretory environment, while the opposite is true for proteins with, for example, an α horseshoe
334 architecture (**Supplementary Table 13**, Figshare links to data in **Online Methods**). Similarly,
335 certain Pfam domains, such as the AAA18-domain (PF13238), are more prominent in enriched
336 fragments than depleted fragments, while many typically cytoplasmic domains such as
337 ribosomal protein domains or the tetratricopeptide repeat (TPR, PF13181) are found
338 exclusively in depleted fragments (Figshare links to data in **Online Methods,** see also
339 **Supplementary Fig. 18**). This illustrates that sequence- and fold-contextual patterns of
340 features still contain much information that is not apparent from averaged parameters.

341

342 *Protein secretability does not correlate with endogenous secretion*

343 Most patterns found to be enriched in secretable fragments are not recapitulated in
344 endogenous secretory proteins. We therefore further evaluated whether the human proteins
345 from which secretable fragments were derived, were enriched in secretory proteins. Since
346 many proteins produced both secretable and depleted fragments in our screens, we
347 considered only those proteins for which no depleted fragments were found. In both *S.*
348 *cerevisiae* and *P. pastoris* screens, the proportion of secretory proteins (i.e., with a signal
349 peptide in the endogenous setting) in this set was not significantly higher than the fraction of
350 secretory proteins in the human proteome (Fisher's one-sided exact test, p=0.9183 (*S.*
351 *cerevisiae*) and p=0.421 (*P. pastoris*), **Table 1**). This disconnect between highly secretable
352 fragments and endogenous secretory proteins likely reflects how features that determine
353 highest efficiency passage through the secretory system are not the most important
354 components of evolutionary selective pressure for secretory proteins.

355

356 **Discussion**

357 Despite the tremendous strides made in the field of recombinant protein production,
358 heterologous expression remains highly unpredictable. A deeper understanding of the intricate
359 ways in which different processes integrate to produce the full set of secretory system proteins,
360 heterologous or not, is only slowly emerging. Although the study of model secretory proteins
361 has led to substantial progress in the field, a more global approach is needed to gain a more
362 profound and comprehensive characterization of the factors that influence secretion.

363 Our SECRiFY method assesses the secretability of proteins on a proteome-wide scale and at
364 domain-sized resolution by yeast. To this end, inspired by developments in the field of massive
365 parallel sequencing library construction and random approaches to protein engineering, we
366 first developed a streamlined method for the construction of a directionally cloned, normalized,
367 and random primed cDNA fragment library. This novel combination of features enabled us to
368 screen the human proteome for secretability at an unprecedented scale and depth compared
369 to previous studies[52,53,19]. By extension, libraries of this type could be valuable alternatives in
370 other applications where fragment screening is beneficial, such as high throughput protein-
371 protein interactomics.

372 In SECRiFY, yeast surface display screening of these libraries is combined with high efficiency
373 cell sorting and deep sequencing to segregate and identify protein fragments that can
374 productively pass through the yeast secretory system. As such, we demonstrated that the

375 secretability of protein fragments across entire proteomes can be verified experimentally in an
376 efficient, systematic, high-throughput and reproducible manner. Although we here used the
377 human proteome as a testcase, our approach is generic and can be used to screen any
378 eukaryotic, or with minor adaptations, even prokaryotic proteome. Already, the databases we
379 have generated in this work constitute by far the largest resources of such yeast-secretable
380 human protein segments, which will be useful in structural studies (where high levels of
381 proteins are required for crystallization), immunological experiments (where recombinant
382 production is needed for vaccine development or antigen discovery), and biochemical
383 characterizations of particular proteins (which also necessitates a minimal protein amount).
384 Remarkably, using both a gradient boosted method based on feature engineering, and an end-
385 to-end trainable convolutional neural network approach, we achieved an AUROC of up to
386 0.790 for *S. cerevisiae* and 0.777 for *P. pastoris.* Practically, this means that secretable and
387 depleted fragments have properties that allow for discrimination, even without prior knowledge
388 of their nature. This unbiased approach confirmed our hypothesis-driven observations that
389 biophysical features relating to secondary structure and flexibility affect secretability.
390 Secretability is thus indeed a learnable feature of protein sequences, a finding that will benefit
391 recombinant protein expression and *de novo* protein design.

392 From a fundamental biology perspective, it is likely that SECRiFY will provide a means to
393 characterize the substrate scope of secretory system processes that regulate secretory protein
394 passage through the eukaryotic secretory system in a proteome-wide manner. This
395 complements existing methods, such as ribosome profiling[54], which deal with protein
396 biogenesis prior to passage through the secretory system.

397 Our screens, and the combined hypothesis-driven and unbiased data mining of the data,
398 uncovered that secretable fragments tend to be less α-helical, more flexible and more
399 intrinsically disordered than fragments without significant display. Chaperones responsible for
400 secretory system quality control are generally poised to recognized mostly exposed
401 hydrophobic stretches[55–57], so conceivably, flexible yet polar and charged fragments would
402 avoid these interactions and quickly progress towards cell wall incorporation. Limited
403 proteolysis measurements also recently confirmed the inverse correlation of in-cell thermal
404 stability with intrinsic disorder, α-helical secondary structure, and aspartate content[58]. Although
405 these measurements involved proteins in their endogenous context in lysed cells, and lack of
406 dataset overlap precluded direct comparison of secretability and thermal denaturation, it would
407 be intriguing to investigate the relationship between secretability and stability. Even though our
408 results show a clear correlation between disorder and secretability, for ordered proteins, quality
409 control mechanisms in the secretory system will generally efficiently remove unfolded or
410 unstable proteins. Indeed, recent limited proteolysis screens of known small protein domains

411 present in PDB or Pfam suggest that structurally defined surface-displayed domains have an
412 overall high stability score[59].

413 Fragment detectability effects may also contribute to the observed enrichment of aspartates
414 and glutamates. Phosphomannose proteins of the yeast cell wall, which confer it a net negative
415 charge, may be repelled by negatively charged displayed fragments, leading to more efficient
416 antibody staining due to increased accessibility. Our observation that more fragments with a
417 high overall positive charge (Lys + Arg content) are not secreted could relate to translational
418 ribosomal stalling at polybasic stretches[60]. Although proline-rich proteins are generally
419 localized in the nucleus or cytoplasm[61], and are associated with ribosome slowdown[62], it is
420 possible that the intrinsic ability of prolines to 'lock' conformations or reduce conformational
421 freedom enhances secretory protein stability and hence, display and secretion. Polyproline or
422 pro-rich stretches are also known motifs for binding to a wide variety of other proteins[29,63,64];
423 as such, the human oxidoreductase ERp57 binds calnexin and calreticulin at their pro-rich
424 motif, and peptidyl-prolyl isomerases often act in complexes associated with multiple
425 chaperones. Prolines are also used as gatekeeper residues against aggregation[65], and
426 perhaps by extension, also against degradation and therefore display.

427 Secretable fragments are not enriched in secretory proteins. SECRiFY detects secretion at the
428 fragment level, potentially causing some features affecting full-length protein secretion to be
429 missed. Nonetheless, the absence of correlation also underlines that SECRiFY assesses
430 secretability, i.e. the capacity to be secreted, rather than actual endogenous cellular
431 localization to the secretory system. Indeed, just like most proteins are only marginally stable,
432 endogenous secretory proteins evolved for function, not secretability.

433 Arguably, our method also has its limitations. In the current SECRiFY setup, secretability was
434 measured in the sequence context of the α mating factor prepro sequence at the N-terminus,
435 and the Sag1 cell wall protein at the C-terminus. While results from our and other labs have
436 indicated that for several single proteins, display efficiency correlates with relative secretion
437 levels, it cannot be excluded that, at least for certain fragments, both leader sequence and the
438 +/- 300 amino acid Sag1 anchor might differentially influence fragment folding, solubility, or
439 stability. In *E. coli*, fusion to large proteins such as SUMO, the *T. harzanium* cellulose binding
440 domain (CBD), or to maltose binding protein (MBP) is an often used strategy to promote
441 'passenger solubilization', although again, effects vary depending on the protein[66,67].
442 Considering the vectorial nature of translation, a C-terminal fusion, as is the case in our setup,
443 is nevertheless generally deemed less perturbing than an N-terminal fusion, although this is
444 not absolute. Sag1 is also a GPI-anchored protein, affecting the entry pathway into the ER[68–
445 70]. Similarly, the prepro leader sequence, with its multi-step processing and preference for

446 posttranslational translocation[71–73], may bias secretability of certain fragments. It remains to be
447 determined whether similar patterns will emerge with different secretory leaders, anchors,
448 promoters, untranslated regions, or growth conditions.

449 Display also imposes limitations on the dynamic range of the method, as there is an upper limit
450 to the number of molecules that the cell wall can accommodate. Generally, this is in the range
451 of about 10^4 molecules per cell[74,75]. Thus, perturbations affecting secretion efficiency in these
452 higher ranges may be missed.

453 In all, with SECRiFY, we here show that our fragment sequence library is proof-of-concept for
454 massively parallel assessment of passage through the secretory system, providing the
455 opportunity the learn which features influence secretability, and what rules sequences must
456 abide by for successful transit through the yeast secretory system. We anticipate that our
457 method and its next-generation derivatives will be of great value in both protein engineering
458 and fundamental studies of the secretory system.

459

460 ## Acknowledgements

461 The authors thank the VIB Nucleomics Core for Illumina sequencing, as well as M. Vuylsteke
462 for statistical insights, K. Vandewalle for help with western blot sample preparation, and Y.
463 Dondelinger for help with plasmid construction. We thank Lennart Martens for freeing up time
464 to discuss the project and to allow D.M, D.T., E.V. and P.R. to spend time on this project away
465 from their main tasks. This work was supported by a Ghent University BOF PhD Fellowship
466 (M.B., H.E.), a PhD Fellowship from the Research Foundation Flanders (FWO) (M.B., H.E.),
467 an FWO research grant (G.0276.13N) (N.C.) and an ERC Consolidator grant no. 616966
468 (N.C.). E.V. is a postdoctoral research fellow of the Research Foundation Flanders. P.R. and
469 W.V. acknowledge support from the Research Foundation Flanders through grant number
470 G.0328.16N. W.D. and J.Z. are funded by the Ghent University Global Campus.

471

472 ## Author Contributions

473 M.B. performed library constructions, methods development and optimization, screenings,
474 sequencing data processing and analyses, and wrote the manuscript. P.R. performed
475 structure-based data-interpretation under the supervision of E.V. and W.V. W.V. ran, analyzed
476 and visualized the biophysical predictions. S.D. and R.B. developed the gradient boosting
477 classifiers, W.D. and J.Z. developed the deep learning classifiers, and together they analysed

the prediction results and feature contributions. D.M. and D.T. developed the website interface and the underlying database. B.V.M. and H.E. provided additional experimental support. N.C. conceived the project, and assisted in experimental design, interpretation, and manuscript writing. All authors read, revised and approved the final version of the manuscript.

Competing Interests

The authors declare that no competing financial interests exist.

References

1. Jan, C. H., Williams, C. C. & Weissman, J. S. Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling. *Science* **346**, 1257521 (2014).

2. Chartron, J. W., Hunt, K. C. L. & Frydman, J. Cotranslational signal-independent SRP preloading during membrane targeting. *Nature* (2016) doi:10.1038/nature19309.

3. Feizi, A., Österlund, T., Petranovic, D., Bordel, S. & Nielsen, J. Genome-Scale Modeling of the Protein Secretory Machinery in Yeast. *PLoS ONE* **8**, e63284 (2013).

4. Barlowe, C. K. & Miller, E. A. Secretory protein biogenesis and traffic in the early secretory pathway. *Genetics* **193**, 383–410 (2013).

5. Braakman, I. & Hebert, D. N. Protein folding in the endoplasmic reticulum. *Cold Spring Harb. Perspect. Biol.* **5**, (2013).

6. Aviram, N. & Schuldiner, M. Embracing the void-how much do we really know about targeting and translocation to the endoplasmic reticulum? *Curr. Opin. Cell Biol.* **29C**, 8–17 (2014).

7. Wolf, D. H. & Schäfer, A. CPY* and the power of yeast genetics in the elucidation of quality control and associated protein degradation of the endoplasmic reticulum. *Curr. Top. Microbiol. Immunol.* **300**, 41–56 (2005).

8. Timms, R. T. *et al.* Genetic dissection of mammalian ERAD through comparative haploid and CRISPR forward genetic screens. *Nat. Commun.* **7**, 11786 (2016).

9. Christianson, J. C. *et al.* Defining human ERAD networks through an integrative mapping strategy. *Nat. Cell Biol.* (2011) doi:10.1038/ncb2383.

10. Wiseman, R. L., Powers, E. T., Buxbaum, J. N., Kelly, J. W. & Balch, W. E. An Adaptable Standard for Protein Export from the Endoplasmic Reticulum. *Cell* **131**, 809–821 (2007).

510 11. Simpson, J. C. *et al.* Genome-wide RNAi screening identifies human proteins with a
511 regulatory function in the early secretory pathway. *Nat. Cell Biol.* **14**, 764–774 (2012).

512 12. Lundberg, E. *et al.* Defining the transcriptome and proteome in three functionally different
513 human cell lines. *Mol. Syst. Biol.* **6**, 450 (2010).

514 13. Geiger, T., Wehner, A., Schaab, C., Cox, J. & Mann, M. Comparative proteomic analysis of
515 eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol. Cell.*
516 *Proteomics MCP* **11**, M111.014050 (2012).

517 14. O'Toole, N., Grabowski, M., Otwinowski, Z., Minor, W. & Cygler, M. The structural
518 genomics experimental pipeline: insights from global target lists. *Proteins* **56**, 201–210 (2004).

519 15. Slabinski, L. *et al.* The challenge of protein structure determination--lessons from structural
520 genomics. *Protein Sci. Publ. Protein Soc.* **16**, 2472–2482 (2007).

521 16. Cornvik, T. *et al.* Colony filtration blot: a new screening method for soluble protein
522 expression in Escherichia coli. *Nat. Methods* **2**, 507–509 (2005).

523 17. Seitz, T. *et al.* Enhancing the stability and solubility of the glucocorticoid receptor ligand-
524 binding domain by high-throughput library screening. *J. Mol. Biol.* **403**, 562–577 (2010).

525 18. Lockard, M. A. *et al.* A high-throughput immobilized bead screen for stable proteins and
526 multi-protein complexes. *Protein Eng. Des. Sel. PEDS* **24**, 565–578 (2011).

527 19. Raeeszadeh-Sarmazdeh, M. *et al.* Identifying Stable Fragments of Arabidopsis thaliana
528 Cellulose Synthase Subunit 3 by Yeast Display. *Biotechnol. J.* e1800353 (2018)
529 doi:10.1002/biot.201800353.

530 20. Martinez Molina, D. *et al.* Engineering membrane protein overproduction in Escherichia coli.
531 *Protein Sci. Publ. Protein Soc.* **17**, 673–680 (2008).

532 21. Luan, C.-H. *et al.* High-throughput expression of C. elegans proteins. *Genome Res.* **14**, 2102–
533 2110 (2004).

534 22. D'Angelo, S. *et al.* Filtering 'genic' open reading frames from genomic DNA samples for
535 advanced annotation. *BMC Genomics* **12 Suppl 1**, S5 (2011).

536 23. Gupta, A. *et al.* A novel helper phage enabling construction of genome-scale ORF-enriched
537 phage display libraries. *PloS One* **8**, e75212 (2013).

538 24. Reich, S. *et al.* Combinatorial Domain Hunting: An effective approach for the identification of
539 soluble protein domains adaptable to high-throughput applications. *Protein Sci. Publ. Protein Soc.* **15**,
540 2356–2365 (2006).

541 25. Yumerefendi, H., Tarendeau, F., Mas, P. J. & Hart, D. J. ESPRIT: an automated, library-based
542 method for mapping and soluble expression of protein domains from challenging targets. *J. Struct.*
543 *Biol.* **172**, 66–74 (2010).

544 26. An, Y., Yumerefendi, H., Mas, P. J., Chesneau, A. & Hart, D. J. ORF-selector ESPRIT: A
545 second generation library screen for soluble protein expression employing precise open reading frame
546 selection. *J. Struct. Biol.* (2011) doi:10.1016/j.jsb.2011.04.004.

547 27. Pedelacq, J.-D. *et al.* Experimental mapping of soluble protein domains using a hierarchical
548 approach. *Nucleic Acids Res.* **39**, e125 (2011).

549 28. Hart, D. J. & Waldo, G. S. Library methods for structural biology of challenging proteins and
550 their complexes. *Curr. Opin. Struct. Biol.* **23**, 403–408 (2013).

551 29. Jadwin, J. A., Ogiue-Ikeda, M. & Machida, K. The application of modular protein domains in
552 proteomics. *FEBS Lett.* **586**, 2586–2596 (2012).

553 30. Boxem, M. *et al.* A Protein Domain-Based Interactome Network for C. elegans Early
554 Embryogenesis. *Cell* **151**, 1633 (2012).

555 31. Prodromou, C., Savva, R. & Driscoll, P. C. DNA fragmentation-based combinatorial
556 approaches to soluble protein expression Part I. Generating DNA fragment libraries. *Drug Discov.*
557 *Today* **12**, 931–938 (2007).

558 32. Zhulidov, P. A. *et al.* Simple cDNA normalization using kamchatka crab duplex-specific
559 nuclease. *Nucleic Acids Res.* **32**, e37 (2004).

560 33. Bogdanov, E. A. *et al.* Normalizing cDNA libraries. *Curr. Protoc. Mol. Biol. Ed. Frederick M*
561 *Ausubel Al* **Chapter 5**, Unit 5.12.1-27 (2010).

562 34. Lyamichev, V., Brow, M. A. & Dahlberg, J. E. Structure-specific endonucleolytic cleavage of
563 nucleic acids by eubacterial DNA polymerases. *Science* **260**, 778–783 (1993).

564 35. Xu, Y. *et al.* Biochemical and mutational studies of the 5'-3' exonuclease of DNA polymerase
565 I of Escherichia coli. *J. Mol. Biol.* **268**, 284–302 (1997).

566 36. Davis, C., Barvish, Z. & Gitelman, I. A method for the construction of equalized directional
567 cDNA libraries from hydrolyzed total RNA. *BMC Genomics* **8**, 363 (2007).

568 37. Howland, S. W., Poh, C.-M. & Rénia, L. Directional, seamless, and restriction enzyme-free
569 construction of random-primed complementary DNA libraries using phosphorothioate-modified
570 primers. *Anal. Biochem.* **416**, 141–143 (2011).

571 38. Mattanovich, D. *et al.* Genome, secretome and glucose transport highlight unique features of
572 the protein production host Pichia pastoris. *Microb. Cell Factories* **8**, 29 (2009).

573 39. Ahmad, M., Hirz, M., Pichler, H. & Schwab, H. Protein expression in Pichia pastoris: recent
574 achievements and perspectives for heterologous protein production. *Appl. Microbiol. Biotechnol.*
575 (2014) doi:10.1007/s00253-014-5732-5.

576 40. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).

577 41. Kliman, R. M., Irving, N. & Santiago, M. Selection conflicts, gene expression, and codon
578 usage trends in yeast. *J. Mol. Evol.* **57**, 98–109 (2003).

579 42. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–741
580 (2003).

581 43. Zur, H. & Tuller, T. Strong association between mRNA folding strength and protein
582 abundance in S. cerevisiae. *EMBO Rep.* **13**, 272–277 (2012).

583 44. Boël, G. *et al.* Codon influence on protein expression in E. coli correlates with mRNA levels.
584 *Nature* **529**, 358–363 (2016).

585 45. Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T. & Vranken, W. F. From protein sequence to
586 dynamics and disorder with DynaMine. *Nat. Commun.* **4**, 2741 (2013).

587 46. Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T. & Vranken, W. F. The DynaMine webserver:
588 predicting protein dynamics from sequence. *Nucleic Acids Res.* **42**, W264-270 (2014).

589 47. Yan, J., Mizianty, M. J., Filipow, P. L., Uversky, V. N. & Kurgan, L. RAPID: fast and
590 accurate sequence-based prediction of intrinsic disorder content on proteomic scale. *Biochim. Biophys.*
591 *Acta* **1834**, 1671–1680 (2013).

592 48. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proc. 22nd ACM*
593 *SIGKDD Int. Conf. Knowl. Discov. Data Min.* 785–794 (2016).

594 49. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

595 50. Raimondi, D., Orlando, G., Pancsa, R., Khan, T. & Vranken, W. F. Exploring the Sequence-
596 based Prediction of Folding Initiation Sites in Proteins. *Sci. Rep.* **7**, 8826 (2017).

597 51. Vanderperre, B. *et al.* Direct Detection of Alternative Open Reading Frames Translation
598 Products in Human Significantly Expands the Proteome. *PLOS ONE* **8**, e70698 (2013).

599 52. Kim, J.-Y., Kim, H. K., Jang, H. J., Kim, E.-K. & Kim, M. K. Optimization of yeast surface-
600 displayed cDNA library screening for low abundance targets. *J. Microbiol. Biotechnol.* **25**, 547–553
601 (2015).

602 53. Bidlingmaier, S., Ha, K., Lee, N.-K., Su, Y. & Liu, B. Proteome-wide identification of novel
603 ceramide-binding proteins by yeast surface cDNA display and deep sequencing. *Mol. Cell. Proteomics*
604 *MCP* (2016) doi:10.1074/mcp.M115.055954.

605 54. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide
606 analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–
607 223 (2009).

608 55. Flynn, G. C., Pohl, J., Flocco, M. T. & Rothman, J. E. Peptide-binding specificity of the
609 molecular chaperone BiP. *Nature* **353**, 726–730 (1991).

610 56. Blond-Elguindi, S. *et al.* Affinity panning of a library of peptides displayed on bacteriophages
611 reveals the binding specificity of BiP. *Cell* **75**, 717–728 (1993).

612 57. Gardner, B. M. & Walter, P. Unfolded Proteins Are Ire1-Activating Ligands That Directly
613 Induce the Unfolded Protein Response. *Science* **333**, 1891–1894 (2011).

614 58. Leuenberger, P. *et al.* Cell-wide analysis of protein thermal unfolding reveals determinants of
615 thermostability. *Science* **355**, eaai7825 (2017).

616 59. Rocklin, G. J. *et al.* Global analysis of protein folding using massively parallel design,
617 synthesis, and testing. *Science* **357**, 168–175 (2017).

618 60. Brandman, O. *et al.* A ribosome-bound quality control complex triggers degradation of
619 nascent peptides and signals translation stress. *Cell* **151**, 1042–1054 (2012).

620 61. Mandal, A., Mandal, S. & Park, M. H. Genome-wide analyses and functional classification of
621 proline repeat-rich proteins: potential role of eIF5A in eukaryotic evolution. *PloS One* **9**, e111800
622 (2014).

623 62. Artieri, C. G. & Fraser, H. B. Accounting for biases in riboprofiling data indicates a major role
624 for proline in stalling translation. *Genome Res.* **24**, 2011–2021 (2014).

625 63. Zarrinpar, A., Bhattacharyya, R. P. & Lim, W. A. The structure and function of proline
626 recognition domains. *Sci. STKE Signal Transduct. Knowl. Environ.* **2003**, RE8 (2003).

627 64. Kay, B. K., Williamson, M. P. & Sudol, M. The importance of being proline: the interaction of
628 proline-rich motifs in signaling proteins with their cognate domains. *FASEB J. Off. Publ. Fed. Am.*
629 *Soc. Exp. Biol.* **14**, 231–241 (2000).

630 65. Reumers, J., Maurer-Stroh, S., Schymkowitz, J. & Rousseau, F. Protein sequences encode
631 safeguards against aggregation. *Hum. Mutat.* **30**, 431–437 (2009).

632 66. LaVallie, E. R. & McCoy, J. M. Gene fusion expression systems in Escherichia coli. *Curr.*
633 *Opin. Biotechnol.* **6**, 501–506 (1995).

634 67. Esposito, D. & Chatterjee, D. K. Enhancement of soluble protein expression through the use
635 of fusion tags. *Curr. Opin. Biotechnol.* **17**, 353–358 (2006).

636 68. Ast, T., Cohen, G. & Schuldiner, M. A Network of Cytosolic Factors Targets SRP-
637 Independent Proteins to the Endoplasmic Reticulum. *Cell* **152**, 1134–1145 (2013).

638 69. Ast, T. & Schuldiner, M. All roads lead to Rome (but some may be harder to travel): SRP-
639 independent translocation into the endoplasmic reticulum. *Crit. Rev. Biochem. Mol. Biol.* (2013)
640 doi:10.3109/10409238.2013.782999.

641 70. Sikorska, N. *et al.* Limited ER quality control for GPI-anchored proteins. *J. Cell Biol.* **213**,
642 693–704 (2016).

643 71. Hansen, W., Garcia, P. D. & Walter, P. In vitro protein translocation across the yeast
644 endoplasmic reticulum: ATP-dependent posttranslational translocation of the prepro-alpha-factor. *Cell*
645 **45**, 397–406 (1986).

646 72. Ng, D. T., Brown, J. D. & Walter, P. Signal sequences specify the targeting route to the
647 endoplasmic reticulum membrane. *J. Cell Biol.* **134**, 269–278 (1996).

648 73. Hessa, T. *et al.* Protein targeting and degradation are coupled for elimination of mislocalized
649 proteins. *Nature* **475**, 394–397 (2011).

650 74. Ryckaert, S., Martens, V., De Vusser, K. & Contreras, R. Development of a S. cerevisiae
651 whole cell biocatalyst for in vitro sialylation of oligosaccharides. *J. Biotechnol.* **119**, 379–388 (2005).

652 75. Chao, G. *et al.* Isolating and engineering human antibodies using yeast surface display. *Nat.*
653 *Protoc.* **1**, 755–768 (2006).

654

Figures

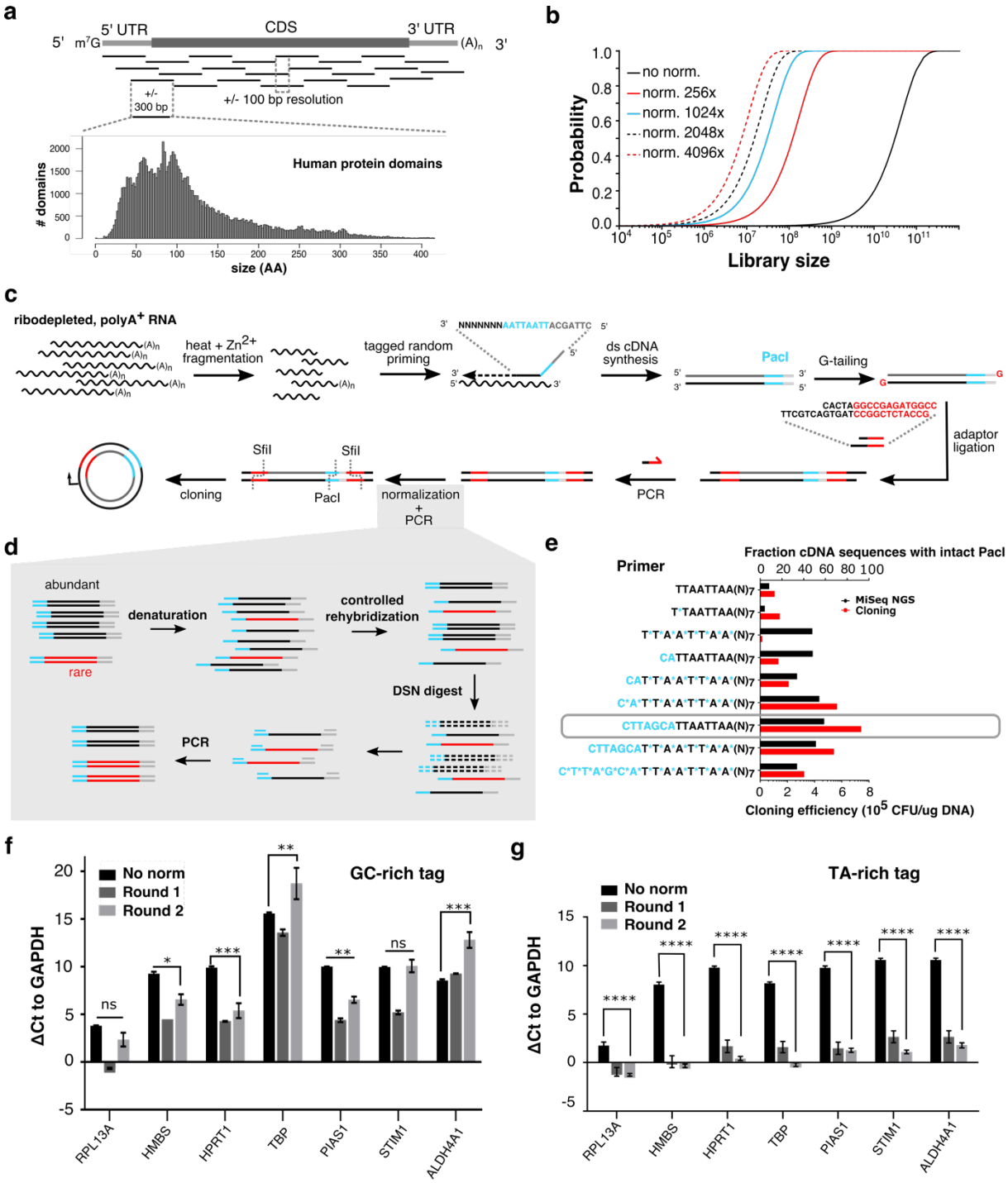

Fig. 1. Capturing protein domains from transcriptomes with directional, normalized fragment libraries. (a) Most protein domains are between 50-150 amino acids (AA) long (lower left, Gene3D (v14.0.1) human protein domains, n= 104,734). Fragmentation of mRNA transcripts to 300 bp fragments should capture a substantial part of the domainome. At a +/- 100 bp resolution, on average 25 fragments would be sufficient to cover a typical transcript.

(**b**). Estimated relationship between library size, at a hypothetical 100 bp resolution, and the probability of sampling any fragment, depending on the efficiency of fragment abundance normalization. **(c)** Fragment libraries are constructed by tagged random priming of fragmented polyA$^+$ RNA, G-tailing, semi-single stranded adaptor ligation, PCR, and duplex-specific nuclease normalization before cloning into the yeast surface display vector. ds= double-stranded. (**d**) Abundant transcripts rehybridize faster than rare ones during kinetically controlled rehybridization after denaturation, and as such, digestion of double-stranded DNA with duplex-specific nuclease (DSN) can be used to normalize fragment abundance. (**e**) Effect of phosphorothioate bonds (blue stars) and buffer sequences (blue nucleotide sequence) on degradation of the PacI sequence in the tag, as measured by deep sequencing (black bars, upper axis) and restriction enzyme/ligase-based cloning into the surface display vector (red bars, bottom axis). The design with buffering bases alone (grey box) was the most effective. Primers are written from 5' to 3'. CFU= colony-forming units, NGS= next-generation sequencing. (**f**, **g**) Abundance differences of various gene fragments compared to *GAPDH*, as ΔCt +/- SEM. All sequence abundances are nearly equalized when using a TA-rich (g), instead of a GC-rich (f) tag in the random primer, with normalization efficiencies up to a +/- 1000-fold (ΔCt of 10) for *HPRT1.* Two-way ANOVA with Tukey post-hoc, ns: non-significant, * $p<0.05$, ** $p<0.01$, *** $p<0.001$, **** $p<0.0001$. n= 9.

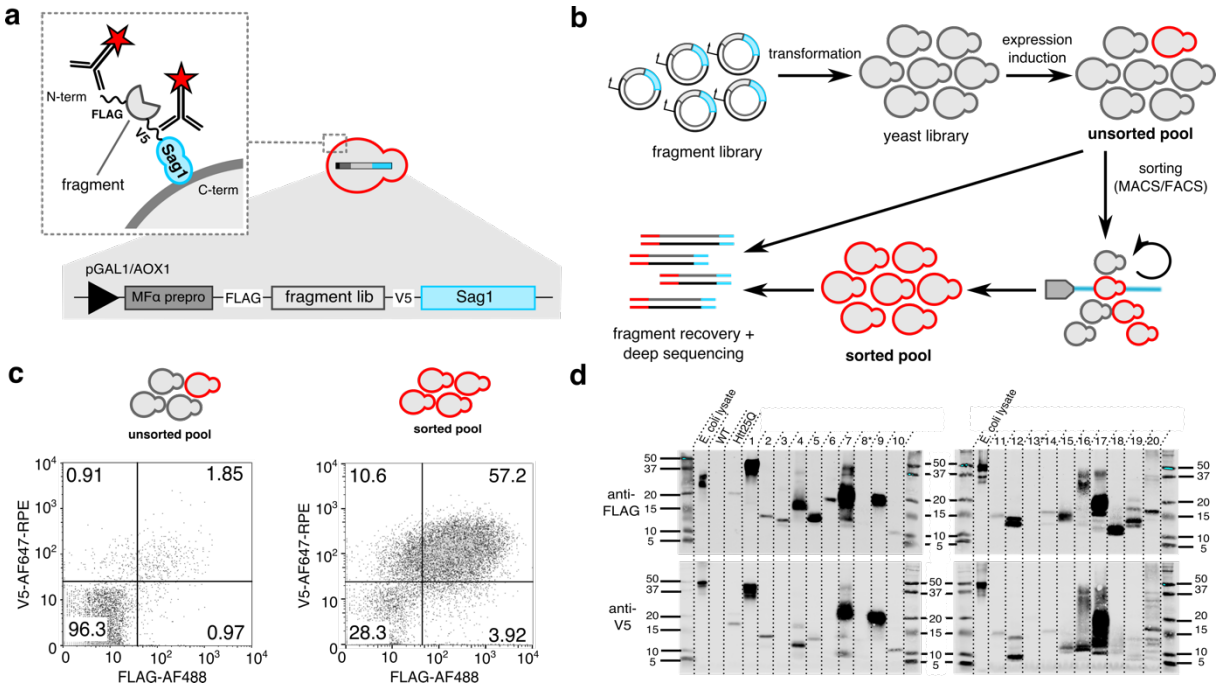

Fig. 2. Screening for secretable protein fragments with the SECRiFY surface display platform. (**a**) Surface display as a proxy for secretion. Libraries are cloned downstream of an inducible promoter (pGal1 for *S. cerevisiae*, and pAOX1 for *P. pastoris*), a secretory leader sequence (MFα prepro), and a FLAG tag; and upstream of a V5 tag and the Sag1 anchor. Productive passage of library polypeptide fragments through the yeast secretory system leads to incorporation into the yeast cell wall, and displaying clones are identified through antibody-based labeling of the epitope tags. (**b**) SECRiFY screening workflow. Fragment libraries are transformed to yeast. After fragment expression induction, displaying FLAG$^+$V5$^+$ clones are sorted in multiple rounds of MACS/FACS. Fragments are identified by PCR recovery and deep sequencing of both sorted and unsorted cell pools. (**c**) Representative flow cytometry plots for SECRiFY screening of the human proteome in *S. cerevisiae*. After 3 rounds of enrichment (MACS/FACS/FACS), the fraction of double-positive (FLAG$^+$V5$^+$) clones increases roughly 30-fold. (**d**) Western blot validation of fragment secretability after SECRiFY screening of the human proteome in *S. cerevisiae*. The majority of human protein fragments from sorted yeast cells can be expressed and secreted into the yeast medium in a Sag1-independent manner. Note that several fragments run as multiple species, likely due to heterogeneous processing and modifications such as O-glycosylation. *E. coli* lysate: antibody positive control, WT: *S. cerevisiae* R1158 medium (neg. control), Htt25Q: medium from *S. cerevisiae* secreting human Htt25Q (pos. control).

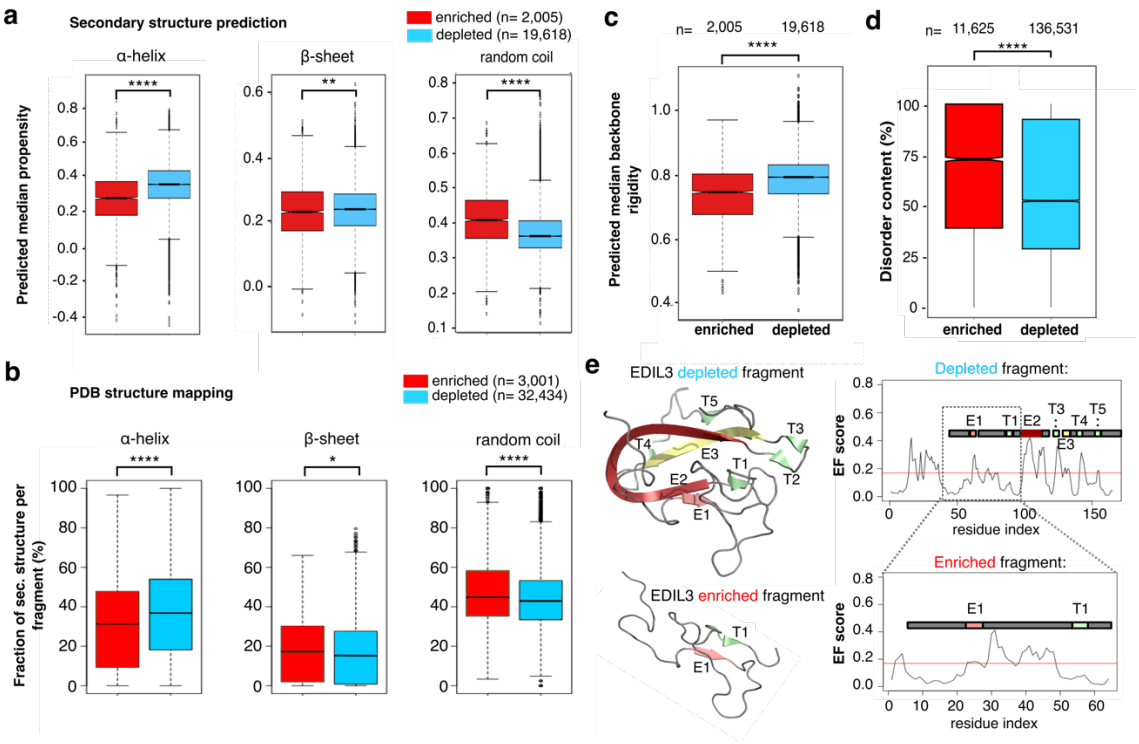

Fig. 3. Patterns in secretable fragments. (a) Dynamine predictions of secondary structure propensity in subsets of consolidated enriched and depleted fragments in *S. cerevisiae.* Enriched fragments have a lower helical content (p= $2.95*10^{-124}$) and a higher random coil (p= $1.99*10^{-127}$) propensity, which is confirmed further by mapping representative fragments to known structures in PDB (α-helix p= $2.35*10^{-8}$, random coil p= $1.24*10^{-4}$) (**b**). Beta sheet differences are not as pronounced (p= $1.26*10^{-3}$ for Dynamine prediction and p= 0.02 for PBD mapping). (**c**) Enriched consolidated fragments are also predicted to be more dynamic than depleted ones (p= $5.08*10^{-118}$). (**d**) Similarly, the predicted disorder content in the total set of enriched fragments is significantly higher than in depleted fragments (p= $<2.2*10^{-16}$). Two-sided Mann-Whitney-Wilcoxon tests in (**a-d**). * p < 0.05, ** p <0.01, **** p<0.0001. (**e**). Two overlapping fragments of the human protein EDIL3 differ in secretability outcome. Early folding (EF) propensity predictions suggest that for the depleted fragment regions E2, T3/E3 and R4 are likely the regions driving folding of the depleted fragment, and lack of these regions in the enriched fragment result in a change in secretability.

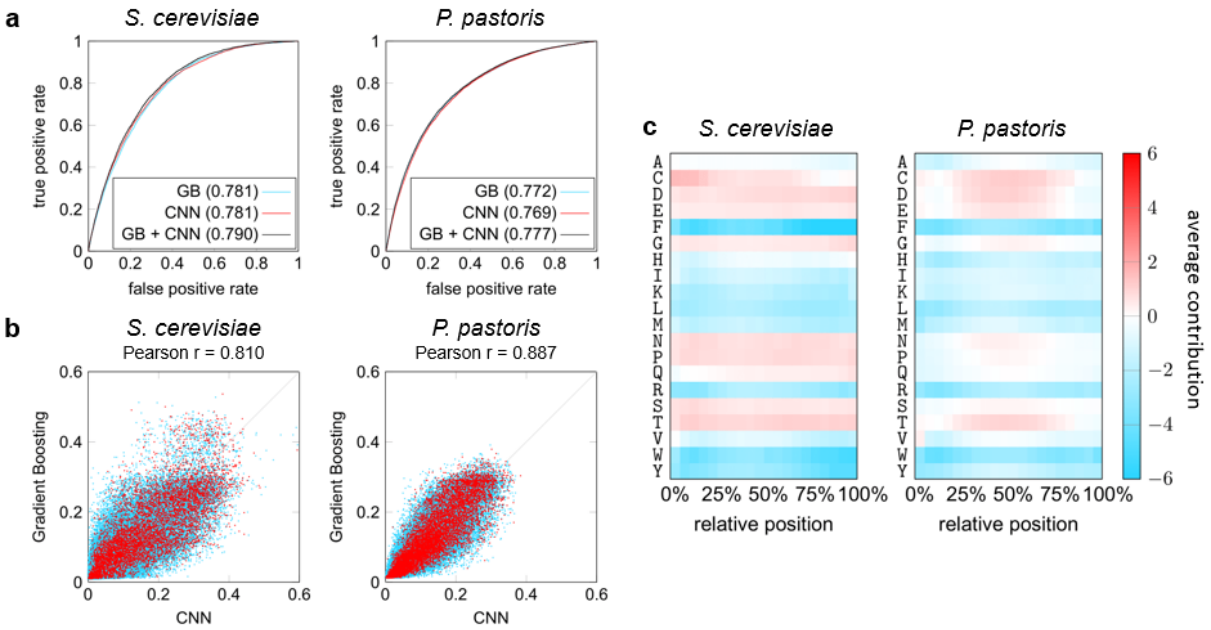

Fig. 4. Machine learning for secretability prediction. (**a**) Evaluation (expressed in AUROC) of the gradient boosting (GB) and convolutional neural network (CNN, one model randomly selected out of the ten trained models) approaches, as well as an ensemble taking the average predictions of both. (**b**) Correlation between the predicted values of the GB and CNN models. Secretable data samples are shown in red, depleted samples in blue. (**c**) Average contribution of individual residues when occurring in different parts of the sequence. For each sequence in a test set, the contribution towards a *positive* prediction (secretable) is calculated for each individual residue. Contributions are then normalized, with absolute values of all contributions in a sequence adding up to 100 on average.

Table 1. Secretory enrichment of secretable proteins.

	with signal peptide	no signal peptide	p-value one-sided Fisher exact test
Human proteome	3574	16548	NA
SECRiFY secretable (*S. cerevisiae*)	17	109	0.9183
SECRiFY secretable (*P. pastoris*)	24	104	0.421