Cryo-EM structure of native human thyroglobulin

Ricardo Adaixo[1], Eva M. Steiner[2], Ricardo D. Righetto[1], Alexander Schmidt[3], Henning Stahlberg[1*] and Nicholas M. I. Taylor[2*]

[1] Center for Cellular Imaging and NanoAnalytics, Biozentrum, University of Basel, Mattenstrasse 26, CH-4058 Basel, Switzerland

[2] Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Blegdamsvej 3B, DK-2200 Copenhagen, Denmark

[3] Proteomics Core Facility, Biozentrum, University of Basel, Klingelbergstrasse 72, CH-4058 Basel, Switzerland

*** Corresponding authors:**

Nicholas M. I. Taylor
Novo Nordisk Foundation Center for Protein Research
University of Copenhagen
Blegdamsvej 3B
DK-2200 Copenhagen N, Denmark
Phone: +45 353 35337
E-mail: nicholas.taylor@cpr.ku.dk

Henning Stahlberg
Center for Cellular Imaging and NanoAnalytics (C-CINA)
Biozentrum, University of Basel
WRO-1058, Mattenstrasse 26
CH-4058 Basel, Switzerland
Phone: +41 61 387 32 62
E-mail: henning.stahlberg@unibas.ch

Running title: Cryo-EM structure of native human thyroglobulin

Keywords: thyroglobulin; cryo-electron microscopy; thyroid; hormonogenesis; cryo-electron microscopy; non-homologous insertions; flexible domains

Abbreviations

ChEL	Cholinesterase-like domain
Cryo-EM	Cryogenic transmission electron microscopy
CTF	Contrast transfer function
DIT	Diiodotyrosination
DSB	Disulfide bond
GlcNAc	N-linked acetylglucosamine
hTg	Human thyroglobulin
MIT	Monoiodotyrosination
MS	Mass spectrometry
NIH	Non-homologous insertions
PTM	Post translational modification
Tg	Thyroglobulin
TH	Thyroid hormone
T3	Triiodothyronine
T4	Thyroxine

Abstract

The thyroglobulin (Tg) protein is essential to thyroid hormone synthesis, playing a vital role in the regulation of metabolism, development and growth. Its structure is conserved among vertebrates. Tg is delivered through the secretory pathway of the thyroid follicular unit to the central colloid depository, where it is iodinated at specific tyrosine sites to form mono- or diiodotyrosine, which combine to produce triiodothyronine (T3) and thyroxine (T4), respectively. Synthesis of these hormones depends on the precise 3D structure of Tg, which has remained unknown despite decades of research. Here, we present the cryo-electron microscopy structure of human thyroglobulin (hTg) to a global resolution of 3.2 Å. The structure provides detailed information on the location of the hTg hormonogenic sites and reveals the position as well as the role of many of its glycosylation sites. Our results offer structural insight into thyroid hormonogenesis and provide a fundamental understanding of clinically relevant hTg mutations, which can improve treatment of thyroid diseases.

63 ## Introduction

64 Thyroglobulin (Tg) is a 660 kDa hyper glycosylated protein expressed in thyrocytes and secreted to the
65 follicular lumen where it accumulates[1]. Dimeric Tg is secreted to the follicular cavity and iodinated to different
66 extents at specific tyrosine residues, a process modulated by the dietary iodine intake[2]. Iodinated Tg is
67 transported to the thyrocyte cytosol by pinocytosis and digested, releasing triiodothyronine (T3) and thyroxine
68 (T4) hormones[3]. Tg is simultaneously a precursor for thyroid hormone (TH) biogenesis and the carrier protein
69 responsible for iodine storage in the follicle colloid. THs are essential to fetal and infant brain development as
70 well as throughout adulthood as metabolism regulators[4]. Mutations in the Tg sequence, or alteration of
71 glycosylation structures, are related to increased risk of thyroid cancer as well as dyshormogenesis associated
72 with goiter[5–7].

73 Analysis of primary sequences of Tg allowed early identification of internal homology domains classified as
74 type 1, type 2 and type 3 repeats[8], as well as a cholinesterase-like domain (ChEL) at the carboxyl end of the
75 protein[9]. Type 1 repeats occur 11 times in the human Tg (hTg) sequence and are homologous to 1a domains
76 found in other proteins with known structure[10,11]. The ChEL domain also has several homologues with
77 structures determined by X-ray diffraction[12–15]. The ChEL domain assists Tg folding, dimerization and
78 secretion processes[16].

79 Despite the extensive biochemical characterization of Tg in the past decades[1,17–22] the three-dimensional
80 structure of Tg remained unknown[1], limiting the understanding of its function. Here, we present the structure
81 of endogenous hTg, determined by cryogenic transmission electron microscopy (cryo-EM). The obtained EM
82 map depicts a dimer with extensive interchain contacts which include but are significantly larger than the
83 ChEL dimer interface. The hTg monomer has 57 disulfide bridges (DSB), which add structural stability and
84 rigidity to most of the protein. However, the extreme N- and C-terminal segments as well as two other regions
85 (the so-called “foot” and “wing”) display a higher degree of flexibility, which is likely related to function.

86 We provide a comprehensive structural description of the endogenous hTg dimer and demonstrate the
87 functional importance of the natural post-translational modification and iodination sites by presenting an
88 atomic model of the nearly complete protein. The native hTg sample is heterogeneous both in composition and
89 conformational states, which likely represents the *in vivo* requirements for hTG function.

90 While this work was in preparation, Coscia et al.[23] described the structure of native, deglycosylated hTG as
91 well as recombinant, non-deglycosylated hTG, identified the putative hormonogenic sites and validated them
92 using an in vitro hormone production assay. Our findings are consistent with those of Coscia et al. but our
93 cryo-EM reconstruction is of the native, non-deglycosylated hTG and of relatively better resolution. We
94 additionally describe novel posttranslational modifications of hTG.

Results

We obtained a homogenous solution of hTg from native sources, by resuspension and gel filtration of the commercially available lyophilate without performing any in vitro iodination. The cryo-EM images of plunge-vitrified hTg solution displayed monodisperse, randomly oriented particles with size and shape consistent with previous observations in negative stain preparations[17]. A cryo-EM map at 3.2 Å nominal resolution was reconstructed allowing atomic modelling of the hTg dimer to around 90% completeness (2,483 modelled residues over 2,748 expected residues per chain). This composite map is the result of a globally-refined consensus map and two maps locally refined around particularly flexible regions, the "wing" and the "foot", as described below **(Fig. 1**, **Supplementary Video 1)**. These peripheral and flexible domains are likely an obstacle to obtain diffraction-quality crystals, which probably prevented the structural determination of hTg in the past. The hTg dimer is approximately 250 Å long by 160 Å wide and 110 Å along the C2 symmetry axis. Each chain is formed by regions I, II and III and a C-terminal ChEL domain **(Fig. 1)**. The interface between monomers buries an area of 31,100 $Å^2$ involving all regions except region II. The hTg structure is annotated similarly to what is reported in the literature[1], however, region I lacks the so called "linker" between repeats 1.4 and 1.5. Three types of cysteine-rich internal homology repeats are present in hTg: type 1, type 2 and type 3. There are 10 type 1 repeats within region I and an 11th in region II. Three type 2 repeats, each bearing 14 to 17 residues, lie between the hinge region and repeat 1.11. Type 3 repeats are located between repeat 1.11 and the ChEL domain.

Type 1 repeats

The first four type 1 repeats in the proximal region I cluster at the N-terminus of each chain and establish extensive contacts with region III and the hinge of the opposing chain, as well as intra chain contacts with repeat 1.5 **(Fig. 1, 2)**. Within repeats 1.1 to 1.4 we observe two non-homologous insertions (NHI), namely on loops 2 and 3 of repeat 1.3. Repeats 1.5 to 1.10 occupy the central core of hTg and this ensemble forms contacts to all other regions on both chains. Additional NHIs are present in loop 1 of repeat 1.5, loop 2 of repeat 1.7 and loop 2 of repeat 1.8 **(Fig. S6)**. Insertions of repeats 1.3 and 1.5 are in close proximity and exposed between the proximal region I and the ChEL domain of the opposing chain. Insertions of repeat 1.8 from both chains lie at the C2 symmetry axis and form a helix bundle providing additional 1,710 $Å^2$ interchain contact surface **(Fig. 3)**. Repeat 1.7 exposes an NHI protruding almost radially to the C2 symmetry axis and forms no additional contacts. We named this protrusion as "foot", and it is flexible as suggested by the diffuse density obtained in the consensus map. Residues 378 to 615 were assigned to the so called "linker region" in previous work while in our structure they form the insertion of repeat 1.5. A consequence of our annotation is that repeat 1.5 encompasses 2 additional cysteines, Cys408 and Cys608, and a total of 4 disulfide bridges, therefore we classify this repeat as type 1c, as opposed to type 1a and type 1b repeats containing 6 and 4 cysteine residues respectively. The remaining NHIs are devoid of additional cysteine residues.

129 *Type 2 repeats*

130 hTg residues 1456 to 1487 comprise three contiguous type 2 repeats flanked by the hinge and repeat 1.11.
131 Each type 2 repeat comprises 2 cysteine residues, all of which are engaged in disulfide bond formation **(Fig.**
132 **4)**. The most N-terminal cysteine of each type 2 repeat establishes a disulfide bond (DSB) with the adjacent N
133 terminal domain while the most C-terminal cysteine establishes a DSB with the adjacent C-terminal domain.
134 Therefore, both repeat 1.11 and hinge region are linked to repeats 2.1 and 2.3 via DSBs contrary to previous
135 hypothesis[8] where all 6 cysteines in type 2 repeats would form DSBs internally.

136 Repeat 2.1 has a small beta-strand between both cysteine residues which is the only secondary structure
137 element found within the set. All type 2 repeats have a shape reminiscent of an arrowhead where the pointy
138 edge of one repeat is embraced by the flat base of the following type 2 repeat.

139 It has been hypothesized that the CXXC motif in type 2 repeats constitutes a thioredoxin box which may be
140 required for Tg multimerization via intermolecular DSBs[24]. We did not investigate this hypothesis. However,
141 all type 2 repeats are extensively solvent exposed and therefore could serve as a potential substrate to
142 thioredoxin.

143 *Type 3 repeats*

144 Region III comprises a total of 5 type 3 repeats linking the ChEL domain to repeat 1.11. These repeats are
145 formed by an alpha helix followed by a three-stranded beta sheet and can be subdivided into type 3a, bearing
146 8 cysteines, and type 3b, bearing 6 cysteines **(Fig. 5)**. The loop connecting the third beta strand to the neighbor
147 domain is longer and apparently disordered. We noticed the previous annotation[1] places the limits of each type
148 3 repeat within secondary structure elements and therefore does not take into consideration the globular nature
149 of individual domains which can be discerned in the structure but, for the sake of consistency, we follow the
150 same annotation.

151 Repeats 3a3 and 3b2 were reasonably well defined in the consensus map (see **Materials and Methods**)
152 however this was not the case for repeats 3a2, 3b1, 3a1 and down to repeat 1.11. Both 2D classes and consensus
153 refinement maps were not well defined in the 3a2 to 1.11 region (which we named "wing region") likely due
154 to increased flexibility. The wing has a "C" shape where the tips seem to act as pivot points, connecting repeats
155 2.3 to 1.11 at the N-terminus and repeats 3a2 to 3b2 at the C-terminus. The reason for the increased flexibility
156 of the wing region is unknown to us.

157 *Mapping of hTg's hormonogenic sites*

158 The post translational iodination of hTG contributes both to thyroid hormonogenesis as well as iodine storage.
159 One hTg monomer contains 66 tyrosines, therefore the hTg dimer represents a huge reservoir for iodination
160 and post-translational modifications. Under sufficient iodide intake 10-15 tyrosine residues become mono- and
161 diiodotyrosines (MIT, DIT), serving as functional hormonogenic units within the hTg structure[3].

162 Hormonogenesis requires a selected pair of donor and acceptor tyrosine residues and four of such sites (A-D
163 sites) were proposed in hTg[21,25] **(Table 1)**. Coupling is proposed between a DIT and another DIT (donor and
164 acceptor) or between a MIT and a DIT (MIT donor and DIT acceptor) to undergo an oxidative quinol-ether
165 coupling reaction to form T4 or T3, respectively[20].

166 To map the iodination status and positions of hTg, we used mass spectrometry (MS) and detected 11 iodinated
167 tyrosine residues that could serve as potential donor and/or acceptor site for TSH production **(Fig. 1, Table 1)**.
168 Out of the detected sites, 10 were MIT, 6 sites MIT and DIT, 3 only MIT and 1 site DIT only. A special role,
169 as highlighted before, plays Y24, the most efficient T4 forming unit, where we detected a thyroxin T4
170 modification as well[21] **(Table 1)**.

171 We did not observe convincing densities for iodine in the cryo-EM map, probably because iodination levels
172 were low, these sites may be particularly sensitive to radiation damage during data collection, the iodination
173 pattern may be inconsistent between different particles, or a combinations of all these factors.

174 Despite the large number of iodination sites, hormonogenesis depends primarily on tyrosine residues near the
175 N-terminus (the A-site with acceptor at Y24, crucial for T4 synthesis) and the C-terminus (the C-site with
176 acceptor at Y2766, specific for T4 and T3 synthesis) respectively. Further proposed hormonogenic sites are
177 site B, with acceptor Y2573[18](iodination detected by MS), site D (acceptor Y131)[25] (no iodination detected by
178 MS) and site E (acceptor Y704)[18](MIT iodination detected and modeled) **(Fig. 6, Table 1)**.

179 In case of the A-site, although we do not directly observe the acceptor Y24 in the map we can deduce its
180 approximate location as the first modeled residue is P30. Furthermore, we revealed Y24 to be clearly mono-
181 and diiodinated and as well showing an additional T4 mass **(Table 1).** Potential donor sites for Y24 have been
182 suggested to be Y234 (donor 1) or Y149 (donor 2)[25]. We detected mono and di-iodination for both donor sites
183 Y149 and Y234 providing further evidence supporting that these residues are potential donors within
184 hormonogenic site A[25]. Taken together with the distance to P30 and flexibility of the N-terminal, this suggests
185 that T4 synthesis can occur within a single hTg monomer.

186 Tyr258 (which is also close to a glycosylation site N484) has been suggested as an alternative donor. In the
187 structure, this residue is still in relatively close position to Pro30 and therefore Y24, however, major
188 conformational changes would be necessary given that the sidechain of Y258 is oriented towards the hTg core.
189 At the B-site, Y2573 is located at the surface of the protein and fairly accessible. The residue is proposed to
190 be an acceptor whereas Y2540 functions as donor **(Fig. 6)**. For both donor and acceptor tyrosines, we observed
191 mono and di-iodination by mass spectrometry and the 6 Å close contact would allow a coupling reaction. In
192 15 Å proximity is as well Y2478 but has not been shown to be iodinated in our sample. Interestingly, T2537,
193 at 3.8Å and 5.6 Å distance to both tyrosines, Y2573 and Y2540, at the B-site, was found to be phosphorylated
194 therefore maybe playing a crucial role in acceleration of hormone production[26]. At about 28 Å distance from
195 this T3 production site, S2441 was detected by mass spectrometry to be sulfonated **(Table 1)**. Sulfonated

196 serines can be involved in various functions including protein assembly and signaling processes and this type
197 of PTM was detected in a cathepsin-C like enzyme from parasites[27].

198 The C-site is not visible in the structure, where T2727 is the last residue traced. It has been shown that Y2766
199 is acceptor to the Y2766 in the neighboring monomer of the hTG dimer[28]. Our structure is perfectly consistent
200 with this observation: the C-α distance between the Thr2708 residues of both monomers is only 20 Å.

201 The D-site, with a captor Y1310, is accessible and exposed. The donor residue has been proposed to be Y108[25].
202 We did not observe iodination of Y1310 nor Y108 and our data does not support those residues as a D-site *in*
203 *vivo*. However, since those residues are surface exposed, *in vitro* iodination was shown to be possible[25].

204 A putative additional site, located around Y704, could also be found in the structure, as it is accessible at the
205 surface of the protein and had monoiodination detected **(Table 1)**. Several donor or acceptor sites are present:
206 Y866 (26 Å distance to Y704), Y883 (20 Å), Y2640 (14 Å), Y2637 (19 Å), but none were detected to be
207 iodinated. However, Y866 and Y883 were found to be iodinated *in vitro* and could therefore be potential
208 donors[18].

209 A wide variety of other iodinated sites (Y785 MIT/DIT, Y1165 MIT, Y1529 MIT, Y2194 MIT/DIT) have
210 been identified and most of these are located on the surface **(Table 1)**. However, it is unclear to what extent
211 they have a role in hormonogenesis as opposed to iodine storage.

212 *Mapping of proteolysis sites*

213 The lifecycle of Tg comprises multiple proteolysis events: the cleavage of the N-terminal 19 residues signal
214 peptide; the N- and C-terminal cleavages which liberate iodopeptides; the limited proteolysis of Tg which
215 releases Tg particles from the colloid agglomerate; and, finally, the digestion of Tg internalized in the
216 thyrocyte[29–31]. Cathepsins, a family of cysteine proteases, perform the mentioned proteolytic attacks on Tg
217 both inside the thyrocyte and in the follicular lumen. The approximate locations of cathepsin proteolysis sites
218 are depicted in **Fig. 3**. Digestion of the extreme N- and C-terminus containing thyroid hormones (TH) is among
219 the earliest proteolysis events experienced by mature Tg[32–34].

220 Insertions of repeats 1.3, 1.5 and 1.8 contain motifs targeted by cathepsins and a sequence of proteolysis events
221 has been suggested[35] involving different proteases from the cathepsin family. Hence, one function of these
222 insertions is to expose proteolysis-prone segments in order to facilitate hTg digestion.

223 Our hTg atomic model lacks the residues between N496-P547 and T1781-N1814 because no clear density was
224 observed in those regions in the EM map. Interestingly, two of the proteolysis sites plotted in **Fig. 3** (major
225 sector marks on positions 500 and 1800) lie within these missing segments, suggesting that our sample was
226 partially digested at these specific locations.

227 *hTG glycosylation and other modifications*

228 The addition of glycan structures to hTg is crucial for protein folding, structure and therefore function,
229 immune-recognition, cell signaling and play a significant role in protein transport and THS production[36].
230 Within the human thyroglobulin monomer 16 N-linked glycosylation sites have been discovered in the mature
231 protein[19,25].

232 We identified 16 N-linked glycosylation sites, per hTG monomer, by mass spectrometry **(Fig. 1, Table 1)**. Out
233 of those, 14 asparagine residues showed an additional density in the map and N-linked glycan structures were
234 modeled. The two glycosylation sites, N198 and N529, are located close to the hormonogenic A-site and within
235 the region of a proteolytic site, respectively and therefore not modelled **(Fig. 1)**. Based on the map an additional
236 glycosylation site at position N2295 was identified and a high mannose was modeled indicating that there
237 might be more than 16 sites in the mature hTg protein **(Table 1)**[25].

238 We present 3 new glycosylation sites at position N198, N1869 and N2122, that have not been described in
239 previous biochemical studies **(Table 1)**[19]. For several of the previously annotated glycosylation sites we did
240 not detect any modification nor did we observe any additional map density; N110, N198, N816 and N1348[19,25].
241 Besides glycosylation, we have also mapped sulfonation, phosphorylation and acetylation. We identified 15
242 acetylation sites, 4 phosphorylation and 4 sulfonation sites **(Table 1)**. No methylations or succinations were
243 detected.

244 The phosphorylation at T2537 (PO_4) and sulfonation Y2540 (SO_3) are within the hormonogenic B-site.
245 Phosphorylation is thought to improve the efficiency of T3 formation[26] and sulfonation of Y24 and the
246 surrounding peptide sequence was shown to be crucial in thyroid hormone synthesis[37,38].

247 Sulfated iodotyrosines (Tyr-S) have a short life before the coupling reaction occurs and it is suggested that
248 after Tyr-S binding to peroxidase where it is iodinated, the sulfate group is removed, releasing an iodophenoxy
249 anion available for coupling with an iodotyrosine donor[38]. For Y2540, the donor residue in the B-site, we
250 detected sulfonation, MIT an DIT representing all 3 states of hormone site preparation and a perfectly prepared
251 hormone formation site **(Fig. 6)**.

252 *Mapping of nonsense and missense mutations*

253 We considered all the nonsense and missense mutations reported previously[1] and plotted these on the hTg
254 structure **(Fig. 7)**. The majority of the mutations causing early termination of hTg translation as well as those
255 causing a change in amino acid identity fall within modelled regions; 31 of 36 nonsense mutations and 84 of
256 90 missense mutations. Interestingly, 4 of the 5 major clusters of the mentioned mutations **(Fig. 7)** overlap
257 with the proteolysis sites depicted in **Fig. 3**, namely those within repeats 1.3, 1.8, 1.10 and to lesser extent the
258 ChEL domain.

259

260 ## Discussion

261 We determined the atomic structure of hTg based on a composite cryo-EM density map at an overall resolution
262 of 3.2 Å for the consensus map. The overall dimeric structure is consistent with previous biochemical
263 experiments, which showed that the cholinesterase domain is necessary and sufficient for hTg dimerization, at
264 least when overexpressed in HEK293 cells[39], but also reveals the participation of regions 1 and 3 in dimer
265 contact formation.

266 The ChEL domains and repeats 1.6 to 1.8 form the core of the dimer and lie in close proximity to the C2
267 symmetry axis while regions 2 and 3 as well as the remaining type 1 repeats occupy more peripheral zones.
268 We modelled 2 of the expected 4 hormonogenic sites, namely site B and site C. Acceptor tyrosines of sites A
269 and C are located at the extreme N and C terminus of the hTg chain and could not be modelled likely due to
270 inherent flexibility of these regions.

271 Our structural characterization also evidences features of hTg that were previously unknown: the
272 environment and possible function of the 4 NHIs present in type 1 repeats; the previously annotated linker
273 region which in fact is a NHI of repeat 1.5; the globular nature of type 3 repeats, which could be annotated
274 differently, and finally the flexible nature of the wing and foot regions.

275 All NHIs are solvent exposed and present peptide motifs recognized by proteases as determined by MS
276 analysis. The NHI of repeat 1.7 is unique in the sense that the remaining NHIs establish contacts to adjacent
277 domains other than the type 1 repeat itself. One speculative hypothesis is that repeat 1.7 NHI could still form
278 inter domain contacts but in the context of hTg multimerization.

279 The hTg atomic structure is decorated with a variety of post translational modifications which were further
280 studied by MS. Importantly, we describe three new glycosylation sites. The density of detected proteolysis
281 sites and the multiple iodination states found for Y24 strongly indicate that our hTg sample was heterogeneous.

282 We expect the hereby-presented structure of native hTG leads to an improved understanding of Tg biology
283 that could be applied in the diagnosis and therapy of thyroid disease, where our model could be valuable e.g.
284 to determine the location of different antibody epitopes and their relation to autoimmune diseases.

285 ## Methods

286 ### *Sample preparation*

287 Human Thyroglobulin (catalog no. T6830; Sigma-Aldrich) was dissolved in gel filtration buffer (25 mM Tris-
288 HCl pH 7.5, 150 mM NaCl, 1x sodium azide) and injected onto a Superdex-200 increase size-exclusion
289 chromatography column connected to an ÄKTA purifier FPLC apparatus (GE Healthcare Bio-Sciences). Peak
290 fractions were pooled and concentrated to 2 mg/mL of protein before plunge freezing.

291 *Cryo-EM sample preparation and data collection*

292 Quantifoil 2/2 400 mesh Cu grids were glow discharged in low pressure air for 30 s. Then 3 µL of concentrated
293 hTg were dispensed to the hydrophilic surface of the grid prior to single side blotting for 2 s and plunge freezing
294 in liquid ethane using a Leica EM GP2 plunger (Leica Microsystems) operating at 20 °C and 80% relative
295 humidity.

296 Frozen grids were imaged using a Titan Krios (Thermo Fisher Scientific) transmission electron microscope
297 operating at 300 kV equipped with a Gatan Quantum-LS energy filter (slit width 20 eV; Gatan Inc.) and a K2
298 Summit direct electron detector (Gatan Inc.). SerialEM[40] was used for automated data collection with 7
299 acquisitions per hole using beam-image shift[41]. Movies were recorded in counting mode with a pixel size of
300 0.64 Å/px at the sample level. Each movie comprised an exposure of 50 $e^-/Å^2$ fractionated into 50 frames over
301 10 s.

302 *Image processing and model building*

303 Movies were preprocessed online in FOCUS[42] using MotionCor2[43] for drift correction and dose weighting and
304 CTFFIND 4.1[44] for contrast transfer function estimation. Out of 8,119 movies acquired, 4,504 had an estimated
305 CTF resolution better than 4 Å and were selected and used for automated particle picking in Gautomatch
306 (Zhang, K., https://www.mrc-lmb.cam.ac.uk/kzhang/Gautomatch/) with a CC threshold of 0.4 using a
307 Gaussian blob as template.

308 Particles were classified in 2D using RELION-3[45] and the best classes were selected for ab-initio map
309 generation and auto-refinement. EMAN2[46] was used to create projections of the refined map. Template-based
310 automated particle picking in Gautomatch was then applied using 20 Å low pass filtered projections as
311 templates. This picking was then applied on 7,266 movies with an estimated CTF resolution better than 6 Å.
312 The new set of particles was pruned by 2D and 3D classification resulting in 37,619 particles being allocated
313 to one class with well-defined features and apparent C2 symmetry. This class was further refined imposing C2
314 symmetry, and particles were corrected for beam-induced motion and CTF refined in RELION-3 **(Fig. S1)**. A
315 consensus map with nominal resolution of 3.3 Å based on the FSC curve at 0.143 criterion[47,48] was obtained
316 after post processing using an automatically estimated B-factor of -45 $Å^2$ **(Fig. S3, S4)**.

317 Particles considered in the consensus map were imported into cryoSPARC v2[49] for localized refinement with
318 the aim of improving quality of the densities in the "wing" and "foot" regions. Masks around these regions
319 were created in UCSF Chimera[50] using a local resolution filtered version of the consensus map as template
320 and the volume segmentation tool. Both regions benefited from the local refinement procedure as the resulting
321 maps display better connectivity and side chain densities, compared to the consensus map in the considered
322 regions **(Fig. S2, S5)**. Interestingly, the best local refinement maps of the wing region were obtained without
323 performing any prior signal subtraction, as judged by visually inspecting the densities.

An atomic model of hTg based on the consensus, "foot" and "wing" maps was built in Coot[51]. The model covers 90% of the amino acid sequence, lacking mainly loops and the N- and C- terminus extensions. Maps were merged using the program phenix.combine_focused_maps and the atomic model was real space refined in PHENIX[52] and validated using MolProbity[53].

Accession codes

The EM map for the complete hTG molecule has been deposited in the EMDB under accession code EMD-12073. Atomic coordinates for hTG have been deposited in the Protein Data Bank under the accession code PDB 7B75.

Acknowledgments

This work was funded by the Swiss National Science Foundation (NCCR TransCure). We thank K. Goldie and L. Kovaczik for support in electron microscopy. Calculations were performed at sciCORE (http://scicore.unibas.ch) scientific computing center at the University of Basel. The Novo Nordisk Foundation Center for Protein Research is supported financially by the Novo Nordisk Foundation (grant NNF14CC0001). N.M.I.T. is a member of the Integrative Structural Biology Cluster (ISBUC) at the University of Copenhagen.

339 ## References

340 1. Citterio, C. E., Targovnik, H. M. & Arvan, P. The role of thyroglobulin in thyroid hormonogenesis. *Nat. Rev.*
341 *Endocrinol.* **15**, 323–338 (2019).

342 2. Izumi, M., Larsen, P. R. Triiodothyronine, thyroxine, and iodine in purified thyroglobulin from patients with
343 Graves' disease. *J. Clin. Invest.* **59**, 1105–1112 (1977).

344 3. Di Jeso, B. & Arvan, P. Thyroglobulin From Molecular and Cellular Biology to Clinical Endocrinology.
345 *Endocr. Rev.* **37**, 2–36 (2016).

346 4. Pharoah, P. O. D., Buttfield, I. H. & Hetzel, B. S. NEUROLOGICAL DAMAGE TO THE FETUS
347 RESULTING FROM SEVERE IODINE DEFICIENCY DURING PREGNANCY. *Lancet* **297**, 308–310
348 (1971).

349 5. Citterio, C. E. *et al.* New insights into thyroglobulin gene: Molecular analysis of seven novel mutations
350 associated with goiter and hypothyroidism. *Mol. Cell. Endocrinol.* **365**, 277–291 (2013).

351 6. Alzahrani, A. S., Baitei, E. Y., Zou, M. & Shi, Y. Metastatic Follicular Thyroid Carcinoma Arising from
352 Congenital Goiter as a Result of a Novel Splice Donor Site Mutation in the Thyroglobulin Gene. *J. Clin.*
353 *Endocrinol. Metab.* **91**, 740–746 (2006).

354 7. Miyoshi, E., Ito, Y. & Miyoshi, Y. Involvement of Aberrant Glycosylation in Thyroid Cancer. *J. Oncol.* **2010**,
355 1–7 (2010).

356 8. Mercken, L., Simons, M.-J., Swillens, S., Massaer, M. & Vassart, G. Primary structure of bovine thyroglobulin
357 deduced from the sequence of its 8,431-base complementary DNA. *Nature* **316**, 647–651 (1985).

358 9. Mori, N., Itoh, N. & Salvaterra, P. M. Evolutionary origin of cholinergic macromolecules and thyroglobulin.
359 *Proc. Natl. Acad. Sci. U. S. A.* **84**, 2813–7 (1987).

360 10. Guncar, G., Pungercic, G., Klemencic, I., Turk, V. & Turk, D. Crystal structure of MHC class II-associated p41
361 Ii fragment bound to cathepsin L reveals the structural basis for differentiation between cathepsins L and S.
362 *EMBO J.* **18**, 793–803 (1999).

363 11. Pavšič, M., Gunčar, G., Djinovi-Carugo, K. & Lenarčič, B. Crystal structure and its bearing towards an
364 understanding of key biological functions of EpCAM. *Nat. Commun.* **5**, 1–10 (2014).

365 12. Sussman, J. L. *et al.* Atomic structure of acetylcholinesterase from Torpedo californica: a prototypic
366 acetylcholine-binding protein. *Science* **253**, 872–9 (1991).

367 13. Cheung, J., Gary, E. N., Shiomi, K. & Rosenberry, T. L. Structures of human acetylcholinesterase bound to
368 dihydrotanshinone I and territrem B show peripheral site flexibility. *ACS Med. Chem. Lett.* **4**, 1091–6 (2013).

369 14. Arena de Souza, V. *et al.* Comparison of the Structure and Activity of Glycosylated and Aglycosylated Human
370 Carboxylesterase 1. *PLoS One* **10**, e0143919 (2015).

371 15. Cheung, J., Mahmood, A., Kalathur, R., Liu, L. & Carlier, P. R. Structure of the G119S Mutant
372 Acetylcholinesterase of the Malaria Vector Anopheles gambiae Reveals Basis of Insecticide Resistance.
373 *Structure* **26**, 130-136.e2 (2018).

374 16. Lee, J., Di Jeso, B. & Arvan, P. The cholinesterase-like domain of thyroglobulin functions as an intramolecular
375 chaperone. *J. Clin. Invest.* **118**, 2950–2958 (2008).

376 17. Kim, K., Chung, J. M., Lee, S. & Jung, H. S. The Effects of Electron Beam Exposure Time on Transmission
377 Electron Microscopy Imaging of Negatively Stained Biological Samples. *Appl. Microsc.* **45**, 150–154 (2015).

378 18. Lamas, L., Anderson, P. C., Fox, J. W. & Dunn, J. T. Consensus sequences for early iodination and
379 hormonogenesis in human thyroglobulin. *J. Biol. Chem.* **264**, 13541–13545 (1989).

380 19. Yang, S. X., Pollock, H. G. & Rawitch, A. B. Glycosylation in human thyroglobulin: Location of the N-linked
381 oligosaccharide units and comparison with bovine thyroglobulin. *Arch. Biochem. Biophys.* **327**, 61–70 (1996).

382 20. Gavaret, J. M., Cahnmann, H. J. & Nunez, J. Thyroid hormone synthesis in thyroglobulin. The mechanism of
383 the coupling reaction. *J. Biol. Chem.* **256**, 9167–9173 (1981).

21. Palumbo, G. Thyroid hormonogenesis. Identification of a sequence containing iodophenyl donor site(s) in calf thyroglobulin. *J. Biol. Chem.* **262**, 17182–17188 (1987).

22. Ohmiya, Y., Hayashi, H., Kondo, T. & Kondo, Y. Location of dehydroalanine residues in the amino acid sequence of bovine thyroglobulin. Identification of 'donor' tyrosine sites for hormonogenesis in thyroglobulin. *J. Biol. Chem.* **265**, 9066–9071 (1990).

23. Coscia, F. *et al.* The structure of human thyroglobulin. *Nature* **578**, 627–630 (2020).

24. Klein, M., Gestmann, I., Berndorfer, U., Schmitz, A. & Herzog, V. The Thioredoxin Boxes of Thyroglobulin: Possible Implications for Intermolecular Disulfide Bond Formation in the Follicle Lumen. *Biol. Chem.* **381**, 593–601 (2000).

25. Berg, G. *et al.* The structure of human thyroglobulin. *Nature* **53**, 113–118 (2020).

26. Citterio, C. E. *et al.* De novo triiodothyronine formation from thyrocytes activated by thyroid-stimulating hormone. *J. Biol. Chem.* **292**, 15434–15444 (2017).

27. Medzihradszky, K. F. *et al.* O-sulfonation of serine and threonine. *Mol. Cell. Proteomics* **3**, 429–443 (2004).

28. Citterio, C. E., Morishita, Y., Dakka, N., Veluswamy, B. & Arvan, P. Relationship between the dimerization of thyroglobulin and its ability to form triiodothyronine. *J. Biol. Chem.* **293**, 4860–4869 (2018).

29. Rousset, B. *et al.* Thyroid hormone residues are released from thyroglobulin with only limited alteration of the thyroglobulin structure. *J. Biol. Chem.* **264**, 12620–6 (1989).

30. Jordans, S. *et al.* Monitoring compartment-specific substrate cleavage by cathepsins B, K, L, and S at physiological pH and redox conditions. *BMC Biochem.* **10**, 23 (2009).

31. Dunn, A. D., Crutchfield, H. E. & Dunn, J. T. Thyroglobulin processing by thyroidal proteases. Major sites of cleavage by cathepsins B, D, and L. *J. Biol. Chem.* **266**, 20198–20204 (1991).

32. Brix, K., Lemansky, P. & Herzog, V. Evidence for extracellularly acting cathepsins mediating thyroid hormone liberation in thyroid epithelial cells. *Endocrinology* (1996) doi:10.1210/endo.137.5.8612537.

33. Brix, K., Linke, M., Tepel, C. & Herzog, V. Cysteine proteinases mediate extracellular prohormone processing in the thyroid. *Biological Chemistry* (2001) doi:10.1515/BC.2001.087.

34. Suban, D. *et al.* Cathepsin C and plasma glutamate carboxypeptidase secreted from Fischer rat thyroid cells liberate thyroxin from the N-terminus of thyroglobulin. *Biochimie* (2012) doi:10.1016/j.biochi.2011.10.018.

35. Friedrichs, B. *et al.* Thyroid functions of mouse cathepsins B, K, and L. *J. Clin. Invest.* **111**, 1733–1745 (2003).

36. Ząbczyńska, M., Kozłowska, K. & Pocheć, E. Glycosylation in the Thyroid Gland: Vital Aspects of Glycoprotein Function in Thyrocyte Physiology and Thyroid Disorders. *Int. J. Mol. Sci.* **19**, 2792 (2018).

37. Venot, N., Nlend, M. C., Cauvi, D. & Chabaud, O. The hormonogenic tyrosine 5 of porcine thyroglobulin is sulfated. *Biochem. Biophys. Res. Commun.* **298**, 193–197 (2002).

38. Nlend, M.-C., Cauvi, D. M., Venot, N. & Chabaud, O. Role of Sulfated Tyrosines of Thyroglobulin in Thyroid Hormonosynthesis. *Endocrinology* **146**, 4834–4843 (2005).

39. Lee, J., Wang, X., Di Jeso, B. & Arvan, P. The cholinesterase-like domain, essential in thyroglobulin trafficking for thyroid hormone synthesis, is required for protein dimerization. *J. Biol. Chem.* **284**, 12752–12761 (2009).

40. Schorb, M., Haberbosch, I., Hagen, W. J. H., Schwab, Y. & Mastronarde, D. N. Software tools for automated transmission electron microscopy. *Nat. Methods* (2019) doi:10.1038/s41592-019-0396-9.

41. Cheng, A. *et al.* High resolution single particle cryo-electron microscopy using beam-image shift. *J. Struct. Biol.* **204**, 270–275 (2018).

42. Biyani, N. *et al.* Focus: The interface between data collection and data processing in cryo-EM. *J. Struct. Biol.* **198**, 124–133 (2017).

43. Zheng, S. Q. *et al.* MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).

44. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J.*

Struct. Biol. **192**, 216–221 (2015).

45. Zivanov, J. *et al.* New tools for automated high-resolution cryo-EM structure determination in RELION-3. *Elife* **7**, (2018).

46. Tang, G. *et al.* EMAN2: An extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157**, 38–46 (2007).

47. Harauz, G. & van Heel, M. Exact filters for general geometry three dimensional reconstruction. *Optik.* **73**, 146–156 (1986).

48. Rosenthal, P. B. & Henderson, R. Optimal Determination of Particle Orientation, Absolute Hand, and Contrast Loss in Single-particle Electron Cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).

49. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).

50. Goddard, T. D. *et al.* UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14 (2018).

51. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. Sect. D* **60**, 2126–2132 (2004).

52. Afonine, V. P. *et al.* Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. Sect. D* **74**, 531–544 (2018).

53. Davis, I. W. *et al.* MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35**, W375–W383 (2007).

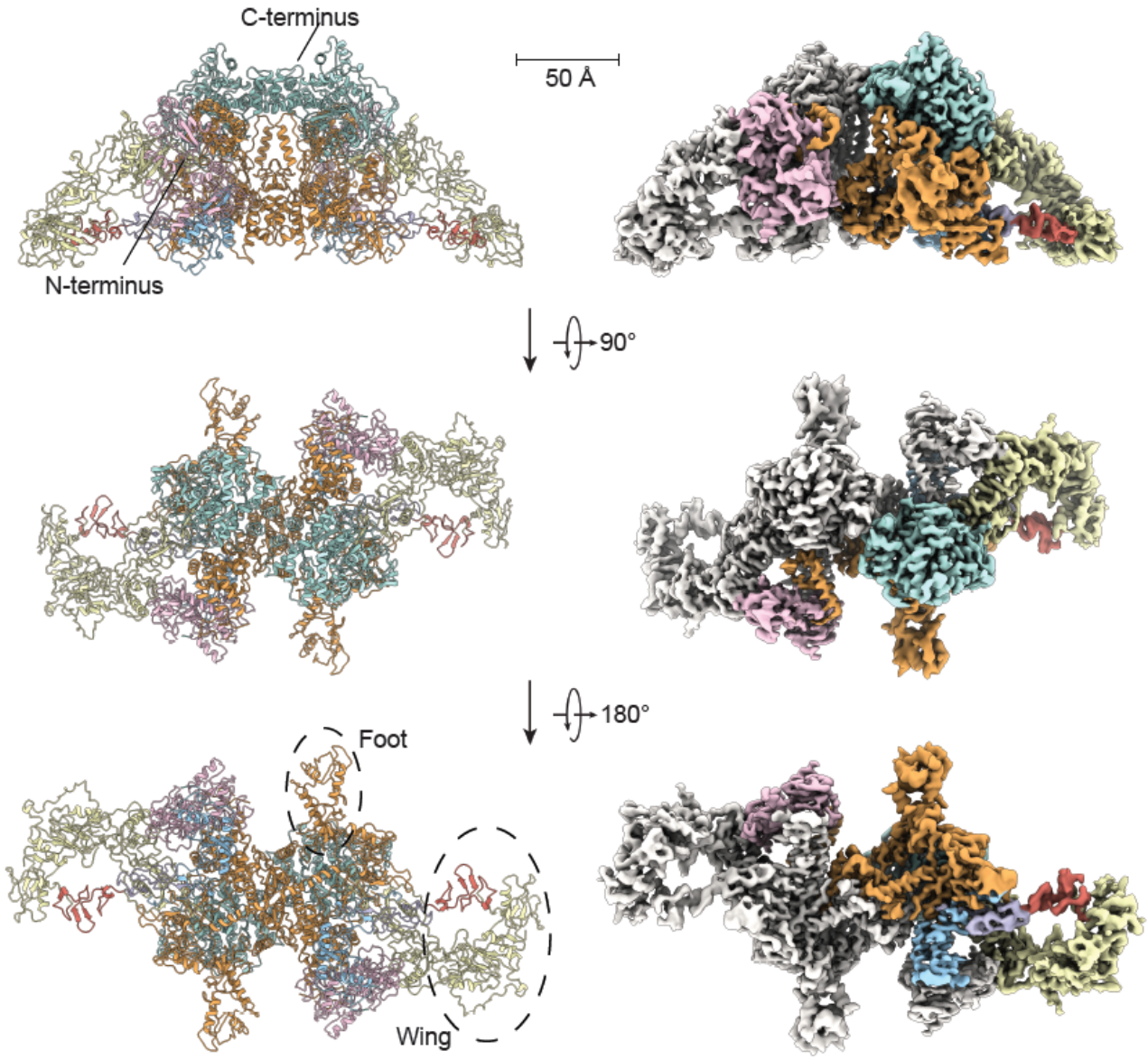
Table 1: List of Native Human Thyroglobulin Modifications Identified by LC-MS

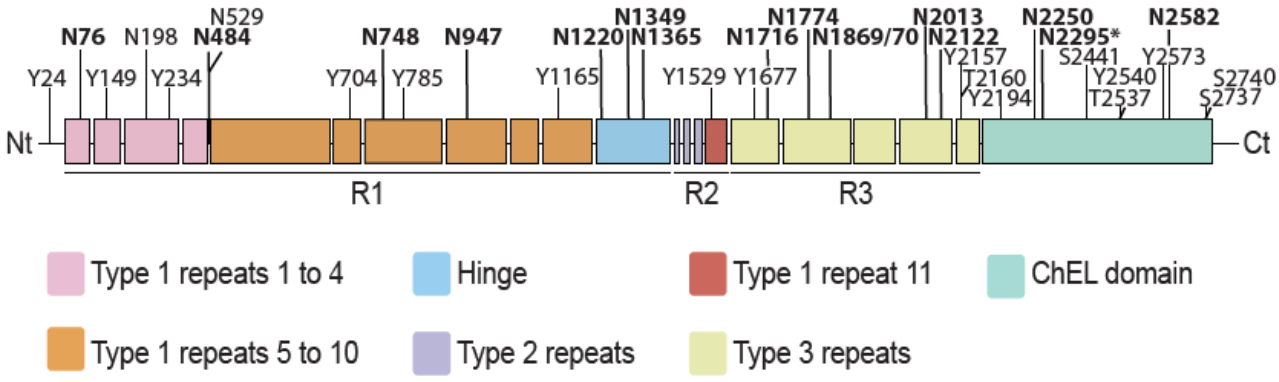
Modification Site	Type of Modifications Identified	Modification known*
Y24	Iodination, di-Iodination, Thyroxin	Yes all
N76	N-Glycosylation	Yes / M
Y149	Iodination, di-Iodination	Yes all
N198	N-Glycosylation	Yes
Y234	Iodination, di-Iodination	No all
N484	N-Glycosylation	Yes / M
N529	N-Glycosylation	Yes
Y704	Iodination	Yes
N748	N-Glycosylation	Yes / M
Y785	Iodination, di-Iodination	Yes, No
N947	N-Glycosylation	Yes / M
Y1165	Iodination	No
N1220	N-Glycosylation	Yes / M
N1348	N-Glycosylation	Yes
N1349	N-Glycosylation	Yes / M
N1365	N-Glycosylation	Yes / M
Y1529	Iodination	No
Y1677	di-Iodination	No
N1716	N-Glycosylation	Yes / M
N1774	N-Glycosylation	Yes / M
N1869	N-Glycosylation	No
N2013	N-Glycosylation	Yes / M
N2122	N-Glycosylation	No / M
Y2157	O-Sulfonation	No
T2160	O-Sulfonation	No
Y2194	Iodination, di-Iodination	No
N2250	N-Glycosylation	Yes / M
N2295	N-Glycosylation	Yes / M
S2441	O-Sulfonation	No
Y2540	Iodination, di-Iodination, O-Sulfonation	No all
Y2573	Iodination, di-Iodination	Yes all
N2582	N-Glycosylation	Yes / M
S2737	Phosphorylation	No

* According to www.uniprot.org (Date of download: 2020/04/20)

M … Modeled Modification

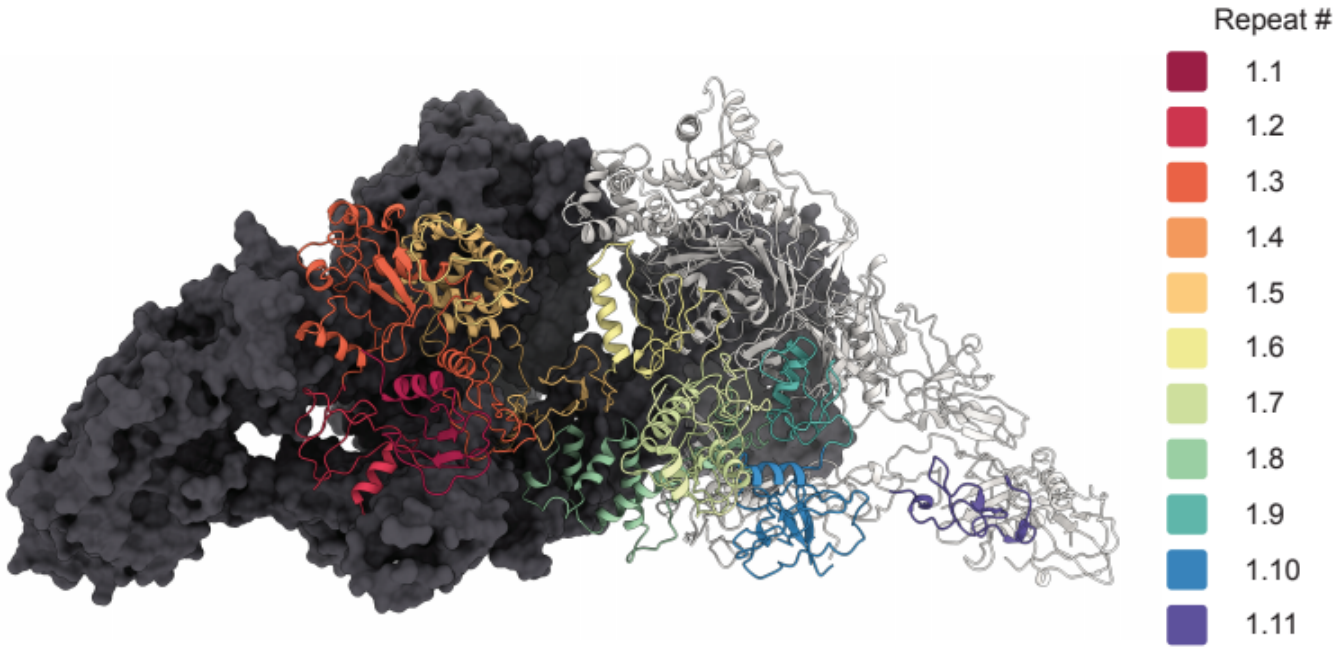
449 **Figures**





450

451 **Figure 1. The cryo-EM map of hTg at 3.2 Å.** (right) Density map of hTg with one
452 monomer in white and corresponding atomic model (left). Map and model are colored as
453 in the bottom linear diagram of a single hTg monomer.



454

455 **Figure 2. Disposition of the different type 1 repeats in the hTg structure.** Surface
456 representation of one hTg monomer in the background colored charcoal. Ribbon
457 representation of the second hTg monomer in the foreground with each type 1 repeat
458 colored as in the scheme on the right.

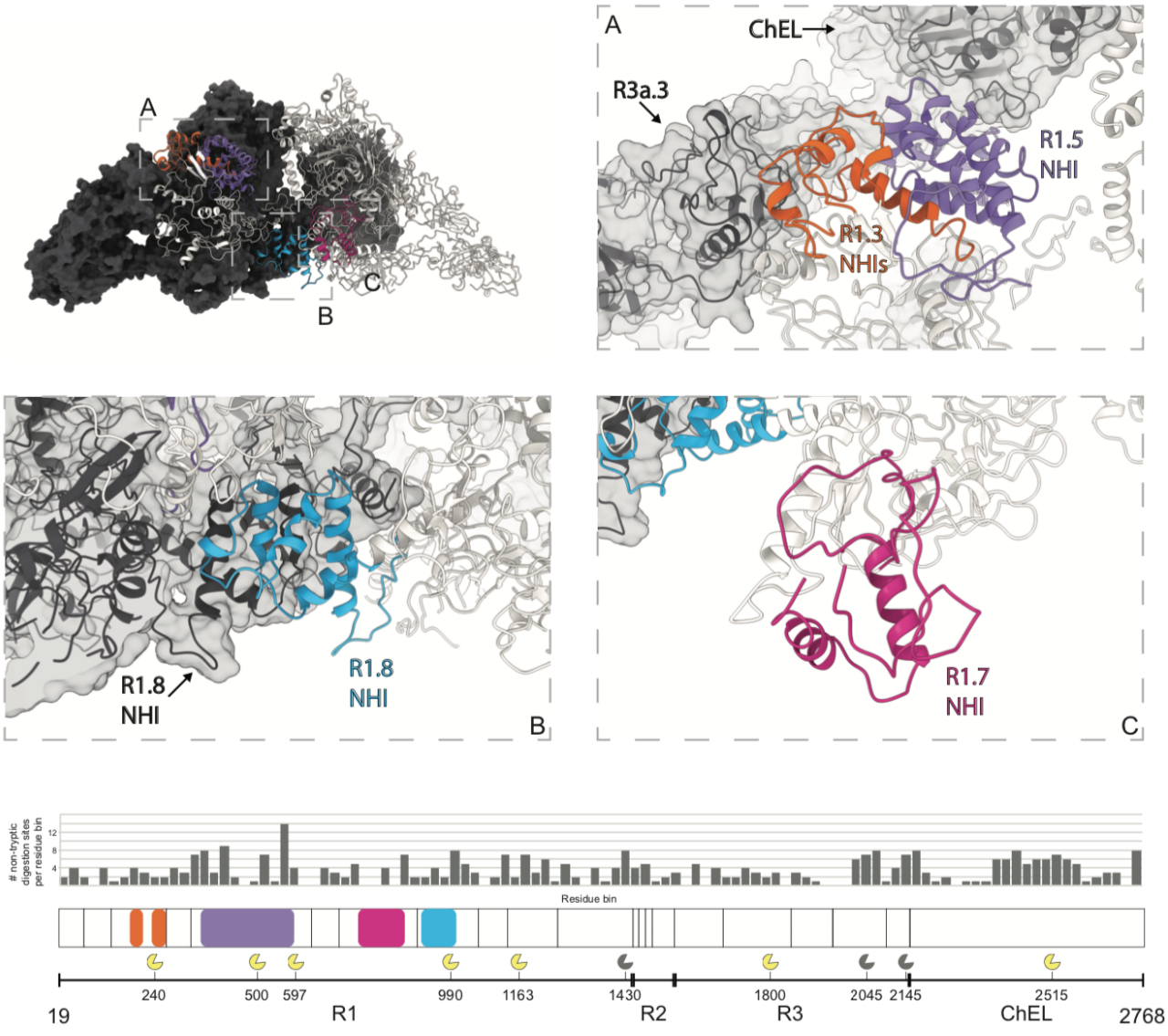
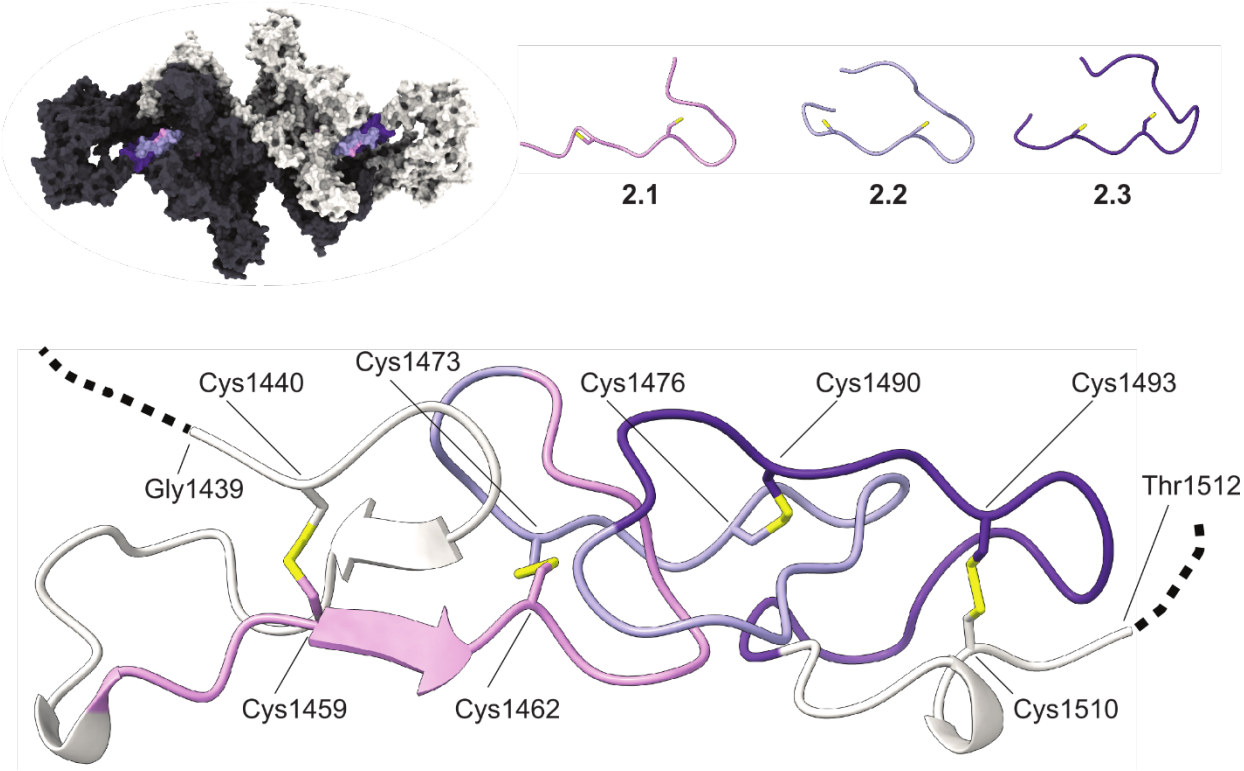


Figure 3. Location of type 1 repeat NHIs and proteolysis clusters. (Top) Surface representation of one hTg monomer (in the background) colored charcoal. Ribbon representation of the second hTg monomer in the foreground with each NHI colored differently: orange – repeat 1.3 NHIs; purple – repeat 1.5 NHI; blue - repeat 1.8 NHI; wine – repeat 1.7 NHI. Detailed clipped views of each insertion are displayed in the dashed boxes; box B view direction is the same as the top left image while boxes A and C were reoriented for better depiction. (Bottom) Histogram of the non-tryptic cleavage sites

467 detected by MS with major sectors depicting the approximate position of previously
468 reported (yellow) and novel (grey) cleavage clusters.



469

470 **Figure 4. Type 2 repeats.** (Top) Ribbon representation of the aligned type 2 repeats.
471 (Bottom) Ribbon representation of the type 2 repeats as disposed in the context of the
472 hTg structure. Cysteine residues represented in sticks.

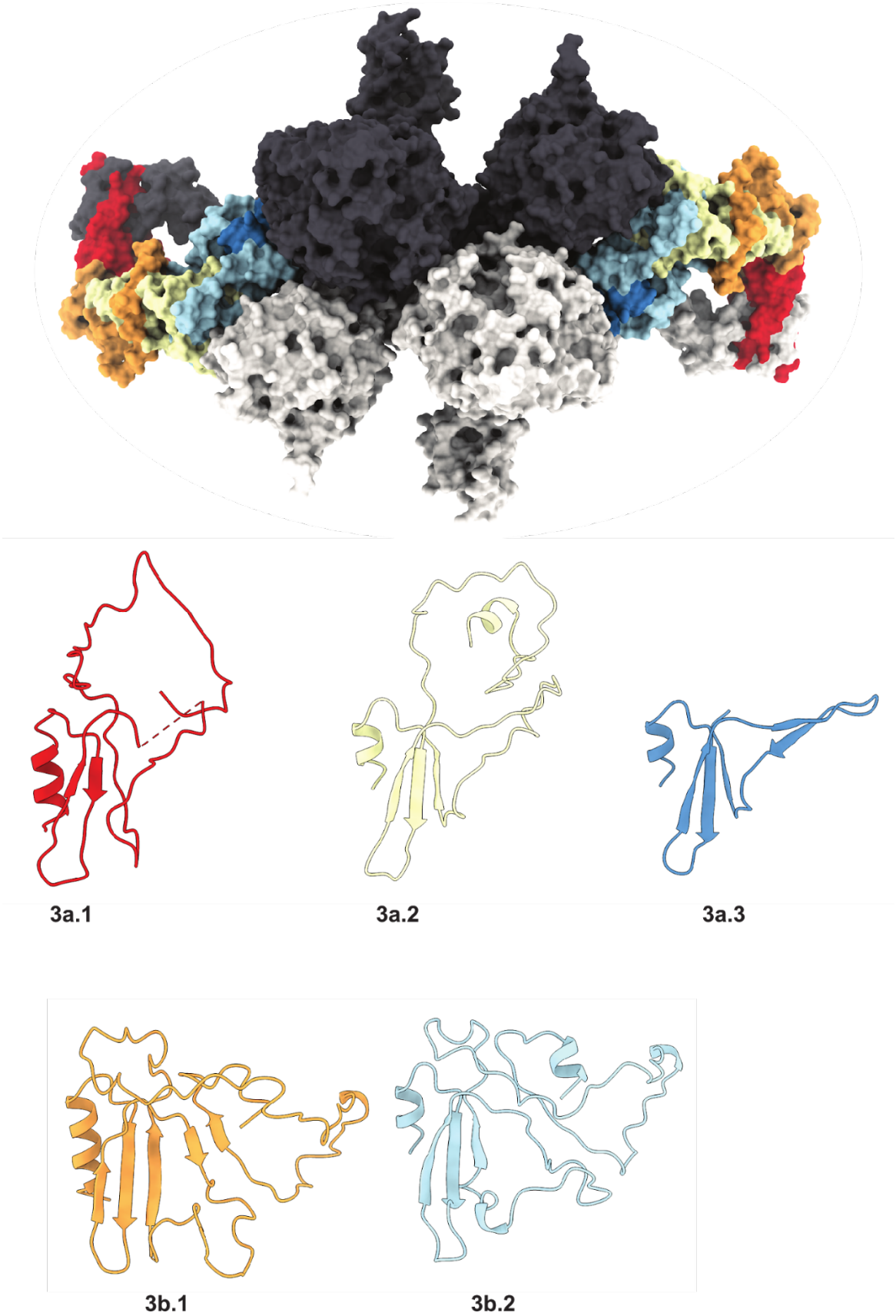


Figure 5. Type 3 repeats. Ribbon representation of the aligned type 3 repeats.

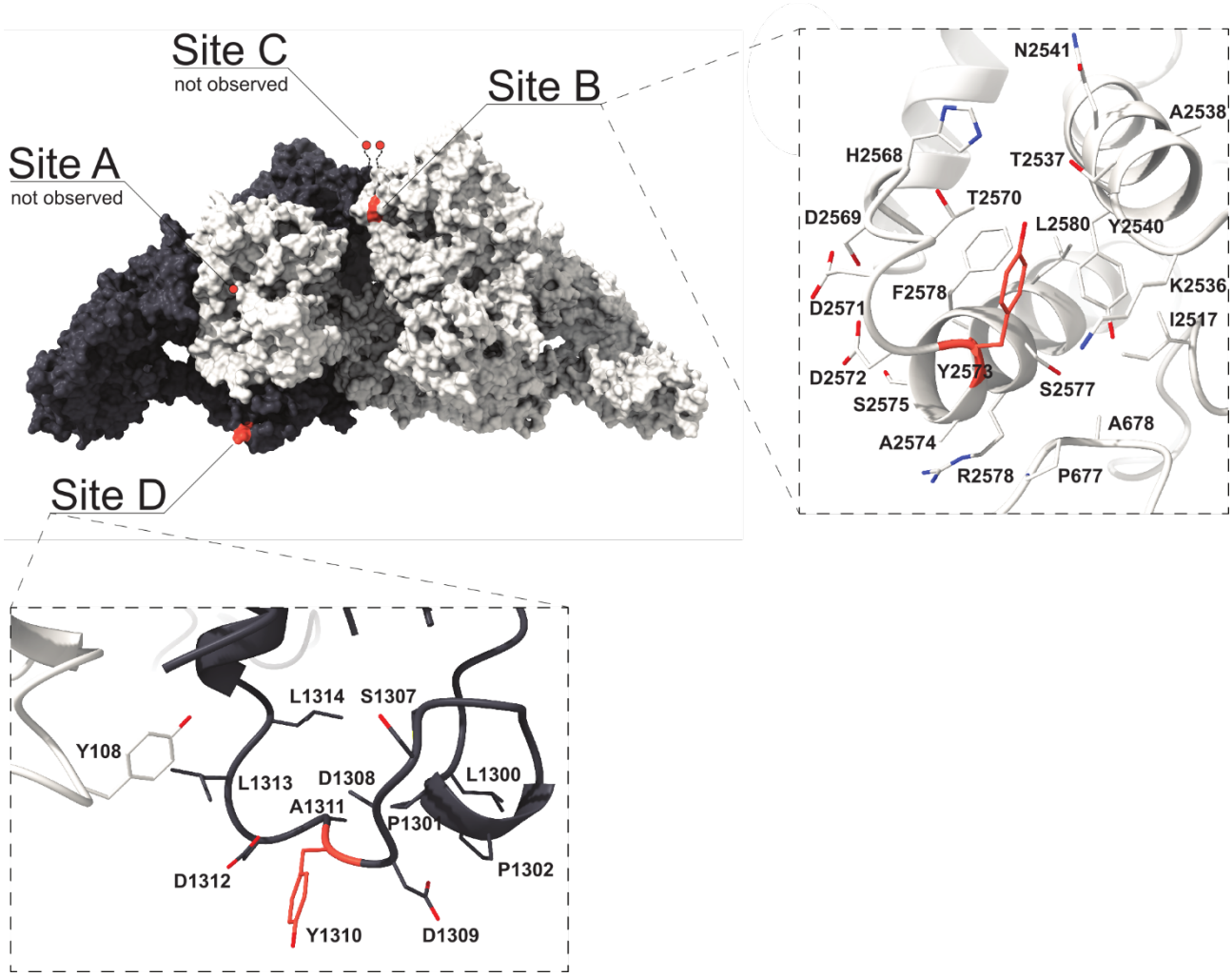


Figure 6. Location of hormonogenic sites in hTg. (Top Left) hTg surface representation with one monomer in charcoal and one monomer in white. Acceptor tyrosine residues in each site (A to D) are marked in orange.

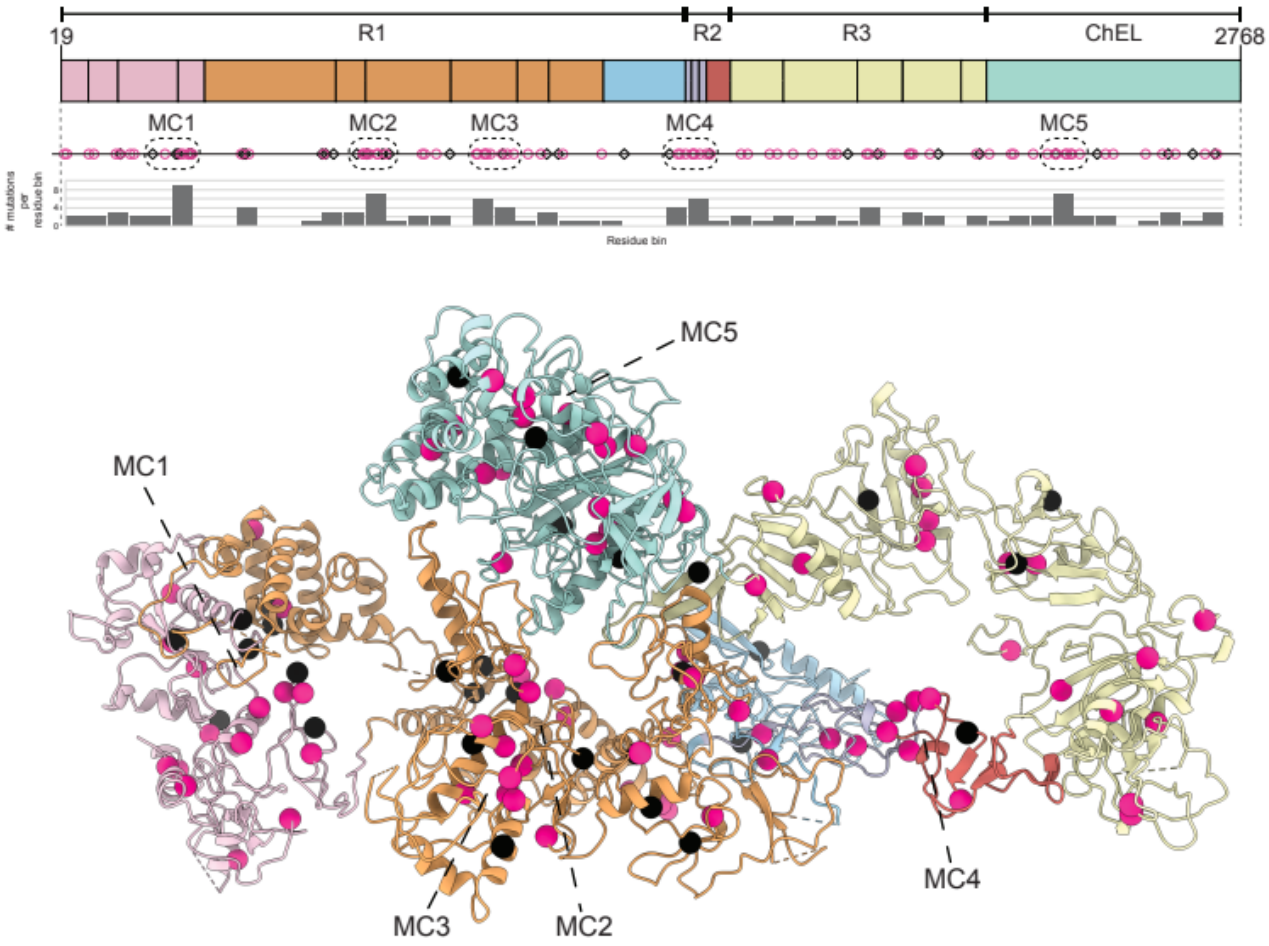


Figure 7. Location of nonsense and missense mutations in hTg. (Top) hTg linear diagram depicting the location of nonsense mutations in black squares and missense mutations in pink; domains color code is the same as **Fig. 1**. Dashed boxes represent the 5 mutation clusters (MC1 to MC5) with the highest density of mutations. (Bottom) Location of the same mutations in the ribbon representation of hTg.