HAP40 orchestrates huntingtin structure for differential interaction with polyglutamine expanded exon 1

Rachel J. Harding*[1], Justin C. Deme[2,3,4], Johannes F. Hevler[5,6], Sem Tamara[5,6], Alexander Lemak[7], Jeffrey P. Cantle[8], Magdalena M. Szewczyk[1], Xiaobing Zuo[9], Peter Loppnau[1], Alma Seitova[1], Ashley Hutchinson[1], Lixin Fan[10], Matthieu Schapira[1,11], Jeffrey B. Carroll[8], Albert J. R. Heck[5,6], Susan M. Lea[2,3,4], Cheryl H. Arrowsmith*[1,7]

[1] Structural Genomics Consortium, University of Toronto, Ontario M5G 1L7, Canada
[2] Sir William Dunn School of Pathology, University of Oxford, Oxford, UK
[3] Central Oxford Structural Molecular Imaging Centre, University of Oxford, South Parks Road, Oxford, OX1 3RE
[4] Center for Structural Biology, Center for Cancer Research, National Cancer Institute, Frederick, MD 21702, USA.
[5] Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute of Pharmaceutical Sciences, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands
[6] Netherlands Proteomics Center, Padualaan 8, 3584 CH Utrecht, The Netherlands
[7] Princess Margaret Cancer Centre and Department of Medical Biophysics, University of Toronto, Toronto, Ontario M5G 1L7, Canada
[8] Behavioral Neuroscience Program, Department of Psychology, Western Washington University, Bellingham, WA, 98225, United States
[9] X-ray Science Division, Argonne National Laboratory, Lemont, Illinois, 60439 USA
[10] Basic Science Program, Frederick National Laboratory for Cancer Research, SAXS Core of NCI, National Institutes of Health, Frederick, Maryland 21701
[11] Department of Pharmacology & Toxicology, University of Toronto, Toronto, Ontario M5S 1A8, Canada

*Corresponding authors - Rachel.Harding@utoronto.ca, Cheryl.Arrowsmith@uhnresearch.ca

Abstract

Huntington's disease results from expansion of a glutamine-coding CAG tract in the huntingtin (HTT) gene, producing an aberrantly functioning form of HTT. Both wildtype and disease-state HTT form a hetero-dimer with HAP40 of unknown functional relevance. We demonstrate *in vivo* that HTT and HAP40 cellular abundance are coupled. Integrating data from a 2.6 Å cryo-electron microscopy structure, cross-linking mass spectrometry, small-angle X-ray scattering, and modeling, we provide a near-atomic-level view of HTT, its molecular interaction surfaces and compacted domain architecture, orchestrated by HAP40. Native mass-spectrometry reveals a remarkably stable hetero-dimer, potentially explaining the cellular inter-dependence of HTT and HAP40. The polyglutamine tract containing N-terminal exon 1 region of HTT is dynamic, but shows greater conformational variety in the mutant than wildtype exon 1. By providing novel insight into the structural consequences of HTT polyglutamine expansion, our data provide a foundation for future functional and drug discovery studies targeting Huntington's disease.

Introduction

The autosomal dominant neurodegenerative disorder Huntington's disease (HD) is caused by the expansion of a CAG repeat tract at the 5' of the *huntingtin* gene above a critical threshold of ~35 repeats [1]. CAG tract expansion corresponds to an expanded polyglutamine tract of the Huntingtin (HTT) protein which functions aberrantly compared to its unexpanded form [2]. Polyglutamine expanded HTT is thought to be responsible for disrupting a wide range of cellular processes including proteostasis [3,4], transcription [5,6], mitochondrial function [7], axonal transport [8] and synaptic function [9]. HD patients experience a range of physical, cognitive and psychological symptoms and longer repeat expansions are associated with earlier disease onset [10]. The prognosis for HD patients is poor, with an average life expectancy of just 18 years from the point of symptom onset and a continuous deterioration of quality of life through this manifest period. There are currently no disease-modifying therapies available to HD patients.

Huntingtin (HTT) is a 3144 amino acid protein comprised of namesake HEAT (Huntingtin, Elongation factor 3, protein phosphatase 2A, TOR1) repeats and is hypothesised to function as a scaffold for larger multi-protein assemblies [11,12]. Many proteomics and interaction studies suggest HTT has an extensive interactome of hundreds of proteins but the only biophysically and structurally validated interactor of HTT is the so-called 40-kDa huntingtin-associated protein HAP40 [13,14], an interaction partner conserved through evolution [15,16]. HAP40 is a TPR domain protein with suggested functions in endocytosis [17–19]. An earlier 4 Å mid-resolution cryo-electron microscopy (cryo-EM) model of HTT in complex with HAP40 reveals that the HEAT subdomains of HTT wrap around HAP40 across a large interaction interface [20]. Biophysical and biochemical analyses comparing purified HTT and HTT-HAP40 samples have revealed that HAP40-bound forms of HTT exhibit reduced aggregation propensity, greater stability and monodispersity as well as conformational homogeneity [20,21]. Consequently, apo HTT is a more difficult sample to work with for structural and biophysical characterisation, and several studies to date have required cross-linking approaches to constrain the HTT molecule to facilitate its analysis, suggesting HTT-HAP40 interactions may stabilize HTT [22,23]. The biological function of the HTT-HAP40 complex however, remains elusive, and it is not clear if the function of this complex differs from apo HTT *in vivo*. It is also not yet understood whether HTT is constitutively bound to HAP40 or whether apo and HAP40-bound forms of HTT perform different functions in the cell.

Current structural information for the full-length HTT molecule sheds little light on the N-terminal exon 1 region of the protein spanning residues 1-90, which contains the critical polyglutamine and polyproline tracts. This region of the protein is unresolved in the HTT-HAP40 cryo-EM model (PDBID: 6EZ8; Guo et al., 2018) and therefore the influence of the tract expansion on HTT structure-function remains unclear. Although many studies have focussed on understanding the effects of polyglutamine expansion on exon 1 in isolation [24–26], there is still very little known about this region in the context of the full-length HTT protein molecule, either in the apo form or in the complex with HAP40. The intrinsically disordered region (IDR), which spans residues 407-665 is subject to a range of post-translational modifications, is postulated to be critical in mediating various protein interactions [21,27,28], and is also unresolved in the cryo-EM structure. Understanding the function of both wildtype and expanded forms of HTT is critical as many potential HD treatments currently under clinical investigation aim to lower HTT expression, using both allele selective or non-selective approaches [29]. Deeper biological insight into the determinants of cellular HTT protein levels, as well as normal and

expanded HTT cellular function would help direct which approaches should be prioritised for long-term patient therapies.

Here, we report *in vivo* studies that show a strong correlation of HTT and HAP40 levels in different genetic backgrounds, providing evidence for the importance of the HTT-HAP40 complex in a physiological setting. Combining the power of multiple complementary structural techniques, we shed light on the missing regions of our high-resolution (2.6 Å) model of HTT-HAP40, including the biologically critical exon 1 region of HTT and the N-terminal region of HAP40. We demonstrate the remarkable stability of the HTT-HAP40 complex, potentially explaining *in vivo* codependence of these two proteins and providing important insight for future drug developments in pursuit of treating HD.

Results

HTT and HAP40 protein levels correlate *in vivo*.

The Huntingtin-associated protein HAP40 co-evolved with HTT [15] and a HAP40 orthologue has been identified in many species, including invertebrates [16]. To investigate the *in vivo* relationship and hypothesised codependency of HTT and HAP40, we analysed the levels of both proteins in liver tissue from different mouse lines using western blot analysis (**Figure 1**). Comparing wildtype (WT) mice, $Htt^{Q111/+}$ Huntington's knock-in mice [30] which express slightly lower levels of HTT [31], and hepatocyte-specific *Htt* knock out mice, a statistically significant correlation was observed for the levels of HTT and HAP40.

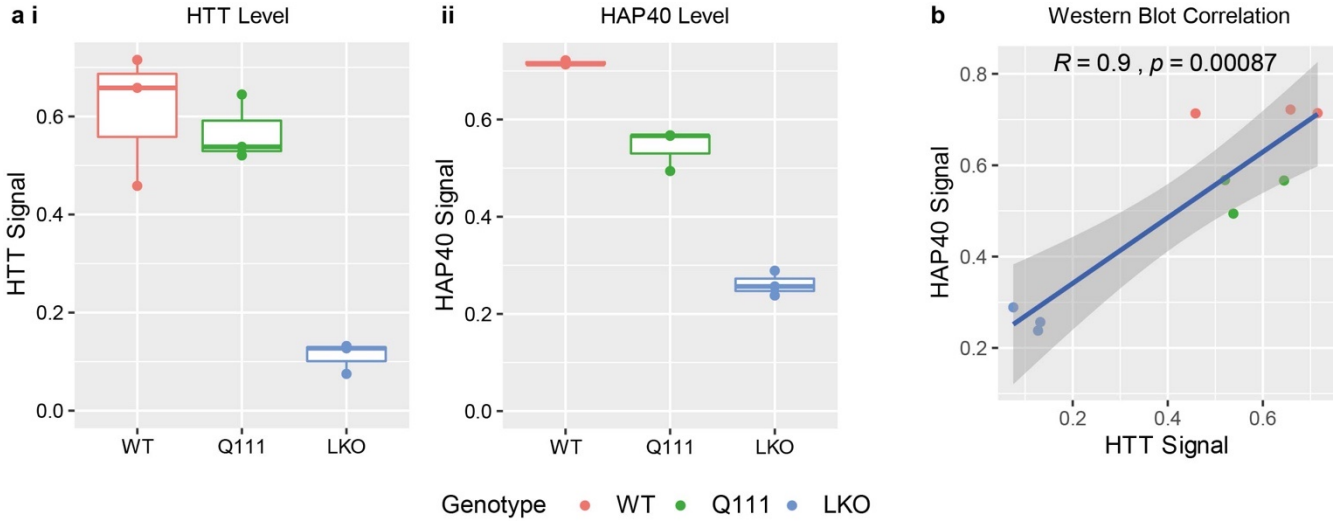


Figure 1. HAP40 levels correlate with the levels of HTT in vivo.
a i HTT and **ii** HAP40 levels were quantified in mouse liver lysates by western blot in wildtype (WT), $Htt^{Q111/+}$ and hepatocyte-specific knockout (LKO) mice. Hepatocytes constitute approximately 80% of liver mass [32] and an approximately 80% reduction in HTT levels is observed in the hepatocyte specific LKO liver tissue as expected. **b** HTT and HAP40 levels correlate in these models with statistical significance.

High-resolution structure of HTT-HAP40 Complex

HTT-HAP40 was expressed in insect cells and purified as previously described [21]. We determined the structure of HTT-HAP40 (PDBID: 6X9O) to a nominal resolution of 2.6 Å using cryo-EM (**Figure 2a**, **Figure 2b** and **Supplementary Figure 1**), improving substantially upon the previously published 4 Å model (PDBID: 6EZ8; Guo et al., 2018) and two recently deposited models (PDBIDs: 7DXJ [3.6 Å] and 7DKK [4.1 Å]; Huang et al., 2021). Similar to all previous models, flexible regions accounting for ~25% of the HTT-HAP40 complex, including exon 1 and the IDR, were not resolved in our high-resolution maps (**Figure 2c**). However, our improved resolution permits more confident positioning of amino acid side chains of the protein structure resolved in the maps and more precise analysis of the different features of the structure.

The overall structure of the complex is similar to the previously published model (PDBID: 6EZ8) with an RMSD of 1.9 across the models when superposed. However, key differences exist between the two models (**Figure 2d**). Two additional C-terminal α-helices in the HTT C-HEAT domain spanning residues 3105-3137 are resolved in our model (all residue numbering based on HTT NCBI reference NP_002102.4 sequence), whereas the resolution of two N-terminal α-helices of HAP40 spanning residues 42-82 is lost. The unmodified native HAP40 C-terminus in our model is able to thread into the centre of the C-HEAT domain (**Figure 2e**). This extended interaction of HAP40 with HTT may be responsible for a small shift we observe of the C-HEAT domain, which pivots ~5° relative to the previous model, reducing the interaction interface of HTT-HAP40 from ~5350 Å^2 to ~4700 Å^2. One potential reason for this difference is that the C-terminus of HAP40 in our construct is unmodified whereas Guo and colleagues used a C-terminal Strep-tag in their expression construct which is unresolved in their model. The differences observed for the HTT and HAP40 interface when comparing our high-resolution structural model (PDBID: 6X9O) and the previous mid-resolution model (PDBID: 6EZ8) indicate that the extensive interaction interface is able to accommodate some variation.

Our high-resolution model enables a comprehensive analysis of the surface-charge features of the HTT-HAP40 complex. The HTT-HAP40 interface is predominantly formed by extensive hydrophobic interactions between the two proteins (**Figure 2f**). Previous analysis of this interface has also highlighted a charge-based interaction between the BRIDGE domain of HTT and the C-terminal region of the HAP40 TPR domain [20]. Interestingly, the N-HEAT domain of HTT has a defined positively charged tract spanning almost 40 Å in length and 5-10 Å in width formed between two stacked HEAT repeats in the N-HEAT solenoid (**Figure 2f arrow**). We also conducted an in-depth sequence conservation analysis of both HTT and HAP40, which we mapped to the high-resolution structure of the complex. Interestingly this revealed surfaces of the protein on the HAP40-exposed face as highly conserved, with extended regions of strict conservation partially spanning the C-HEAT domain, BRIDGE and N-HEAT (**Figure 2g**). However, the opposite face is less conserved, whilst the HTT-HAP40 interface is moderately conserved for both HTT and HAP40. The HTT-HAP40 model was searched for ligand-able pockets which were assessed for druggability according to various factors, including their buriedness, hydrophobicity and volume. One of the most promising pockets, which is predicted to be ligand-able, lies at the HTT-HAP40 interface and is lined by residues from the N-terminal region of the HAP40 TPR domain as well as the HTT N-HEAT domain (**Figure 2h, Supplementary Table 2**). The high resolution of our HTT-HAP40 model provides a foundation for virtual screening of such pockets and other structure-based drug-discovery efforts towards the identification of HTT ligands.

Our 2.6 Å structure is of sufficient resolution to allow the identification of post-translational modifications (PTMs). However, no PTMs were observed for any of the resolved residues in the HTT-HAP40 complex. Native mass spectrometry (MS) analysis, on the other hand, revealed the high purity of our HTT-HAP40 samples, albeit that a small mass difference (compared to the theoretical mass) was observed, consistent with the presence of a few PTMs (**Supplementary Figure 2a**). Further analysis of the HTT-HAP40 complex upon Caspase6 digestion revealed these PTMs to be primarily phosphorylations (at least two), which could be mapped to the regions spanning 586-2647 and 2647-3144 of the HTT sequence (**Supplementary Figure 2b, c** and **d**). Based on the cumulative evidence from the MS data, these modifications reside within the two flexible portions of HTT not resolved in our cryo-EM maps. Although many studies have identified numerous different sites and possible PTMs of the HTT protein [21,27,28,34], these approaches have so far been qualitative and do not give us a good understanding of the key proteoforms the Huntington's disease community is studying in either *in vitro* or *in vivo* models. Our quantitative top- and middle-down MS approaches suggest many post-translational modifications are in fact only present at very low abundance, at least in our insect cell expressed samples.

We attempted to separately purify HTT and HAP40 for comparison to the complex. As reported by Guo and colleagues [20], we were also unable to express recombinant HAP40 alone, although it is readily expressed in the presence of HTT, a trend that parallels our *in vivo* observations. In the absence of HAP40, we and others have shown that recombinant HTT self-associates and is conformationally heterogenous *in vitro* [21,22,34]. Cryo-EM analysis of our apo HTT samples yielded a 12 Å resolution envelope (**Figure 3a** and **b**). Despite the low resolution of this envelope, it is possible to identify the N-HEAT domain, with its central cavity, as well as the C-HEAT domain. The HTT portion of our HTT-HAP40 model can be fitted into this envelope. Comparison of this envelope with the previously reported apo HTT cryo-EM envelopes that were stabilized by cross-linking (EMD4937 and EMD10793; [22] shows a less collapsed arrangement of the HTT subdomains. The difference in resolution between apo HTT and HTT-HAP40 samples observed by cryo-EM analysis emphasizes the importance of HAP40 in stabilising the HTT protein and constraining the HEAT repeat subdomains into a more rigid conformation, further supporting the idea that this is a critical interaction for modulating HTT structure and function.

Native top-down MS uses gas-phase activation to dissociate protein complexes enabling identification of complex composition and subunit stoichiometries. The most commonly used activation method using collisions with neutral gas molecules typically results in dissociation of a non-covalent complex into constituent subunits. Interestingly, our native top-down MS analysis of the intact HTT-HAP40 complex (**Figure 4a** and **b**) primarily resulted in backbone fragmentation of HTT, eliminating both N- and C-terminal fragments (**Figure 4c-g**). Remarkably, the vast majority of concomitantly formed high-mass dissociation products retained HAP40 (**Figure 4f**), suggesting that the extensive hydrophobic interaction interface we observe in our high-resolution model keeps the HTT-HAP40 complex exceptionally stable. Similarly, gas-phase activation of Caspase6-treated HTT-HAP40 revealed that HAP40 remained intact and bound to HTT even at the highest activation energies, whereas the N- and C-terminal fragments of HTT produced upon digestion were readily dissociating from the complex (**Supplementary Figure 2c**).

The recombinant samples of HTT-HAP40 were found to be highly monodisperse (**Figure 4b**), displaying optimal biophysical properties (see also **Supplementary Figure 3a**). Systematically screening the stability of the HTT-HAP40 complex using a differential scanning fluorimetry assay indicates the complex is highly stable under a broad range of buffer, pH and salt conditions (**Supplementary Figure 3b** and **c**). Destabilisation of the complex was only observed at low pH (**Figure 4h**). Similarly, the interaction between HTT and HAP40 is retained upon mild proteolysis of the complex (**Figure 4i**, all data in **Supplementary Figure 3d**). Following Caspase-6 treatment, the HTT-HAP40 complex remains associated under native conditions, although HTT cleavage products are observed under denaturing conditions [35]. Taken together, our studies reveal the high stability of the HTT-HAP40 complex with resistance to dissociation by native top-down MS, or proteolytic cleavage in solution. These data further support the high codependence of HTT and HAP40 protein levels in animal models and possibly HD patients.

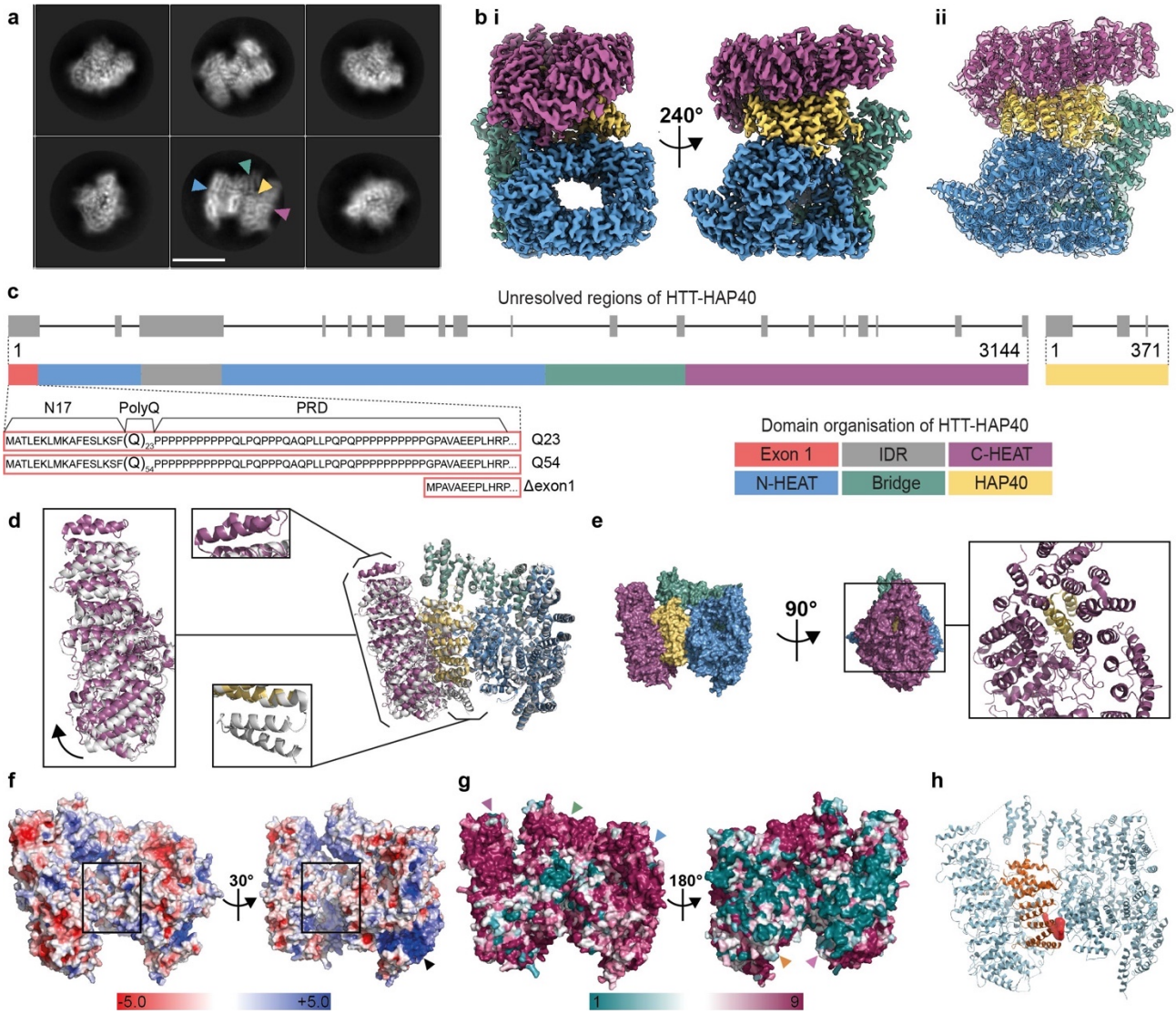


Figure 2. HAP40 stabilises the structure of HTT via extensive interactions across all HEAT repeat subdomains.

a Representative cryo-EM 2D class averages of HTT-HAP40. Scale bar (white) is 90 Å. Blue and purple arrowheads denote N- and C-HEAT domains of HTT, respectively. Green and yellow arrowheads denote bridge domain of HTT and HAP40, respectively. **b** Cryo-EM volume of HTT-HAP40 resolved to 2.6 Å with **i** HTT N-HEAT in blue, bridge domain in green, C-HEAT in purple and HAP40 in yellow or **ii** map shown with HTT-HAP40 modeled in using the same domain colour convention. **c** Domain organisation of HTT mapped to linear sequence. Unresolved regions of the structure are in grey and the three different constructs used in this study are detailed comprising wildtype (23 glutamines; Q23), mutant (54 glutamines; Q54), or HTT with exon 1 partially deleted (Δexon 1; comprising residues 80-3144). **d** Superposition of our model (PDBID: 6X9O – same domain colour convention as before) and the previous model (PDBID: 6EZ8 – all grey) with alignment calculated over N-HEAT and bridge domains. Additional α-helices observed in either of the models are indicated with boxes, C-HEAT domain shift is shown with an arrow. **e** Surface representation of HTT and HAP40 (same domain colour convention as before) in front and side views, rotated 90°, with additional panel (right) showing same side view of the complex in cartoon format. **f** Electrostatic surface representation of HTT with HAP40 removed from the structure. Positively charged regions are shown in blue, neutral (hydrophobic) regions in white and negatively charged regions in red. The positively charged tract in the N-HEAT domain is indicated with a black arrowhead. Hydrophobic HTT surface which binds HAP40, is indicated with hollow black boxes. **g** Surface representation of HTT-HAP40 complex, coloured according to Consurf conservation scores: from teal for the least conserved residues (1), to maroon for the most conserved residues (9). Conserved surfaces for C-HEAT, bridge and N-HEAT domains are indicated with purple, green and blue arrowheads respectively. Variable N-HEAT and C-HEAT surfaces are indicated with orange and pink arrowheads respectively. **h** HTT (pale blue)-HAP40 (orange) complex in cartoon with pocket predicted to be druggable shown as red volume.

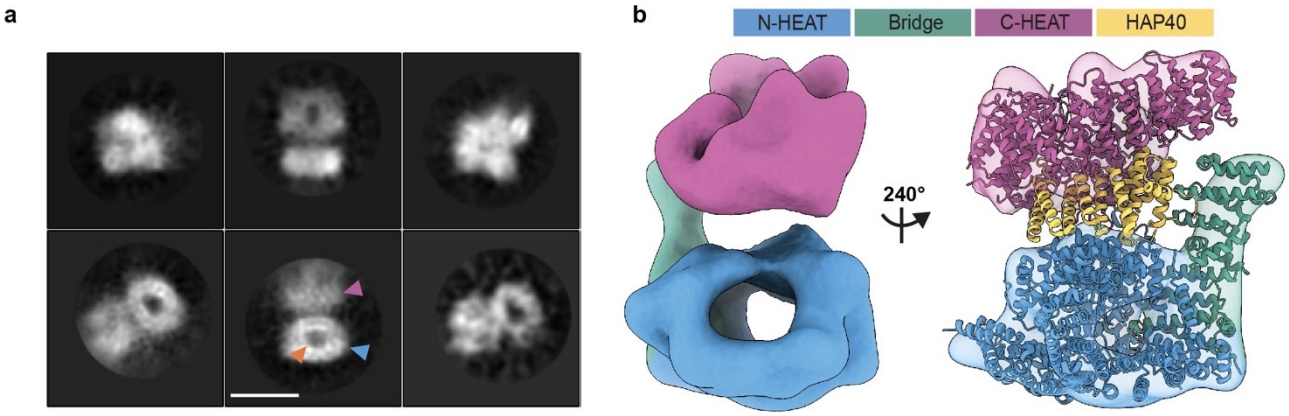


Figure 3. HTT HEAT domains are conformationally flexible in the absence of HAP40.

a Representative cryo-EM 2D class averages of HTT Q23. Scale bar (white) shown in the bottom middle panel is 90 Å. Blue arrowhead denotes the N-HEAT domain in which its central cavity (orange arrowhead) is more clearly defined. Purple arrowhead denotes the less well-defined C-HEAT domain, perhaps due to conformational flexibility relative to the N-HEAT. **b** Cryo-EM volume of HTT resolved to ~12 Å shown with model of HTT-HAP40 (PDBID: 6X9O) fit to the map. Regions of map and model are displayed with N-HEAT in blue, bridge domain in green, C-HEAT in purple and HAP40 in yellow.

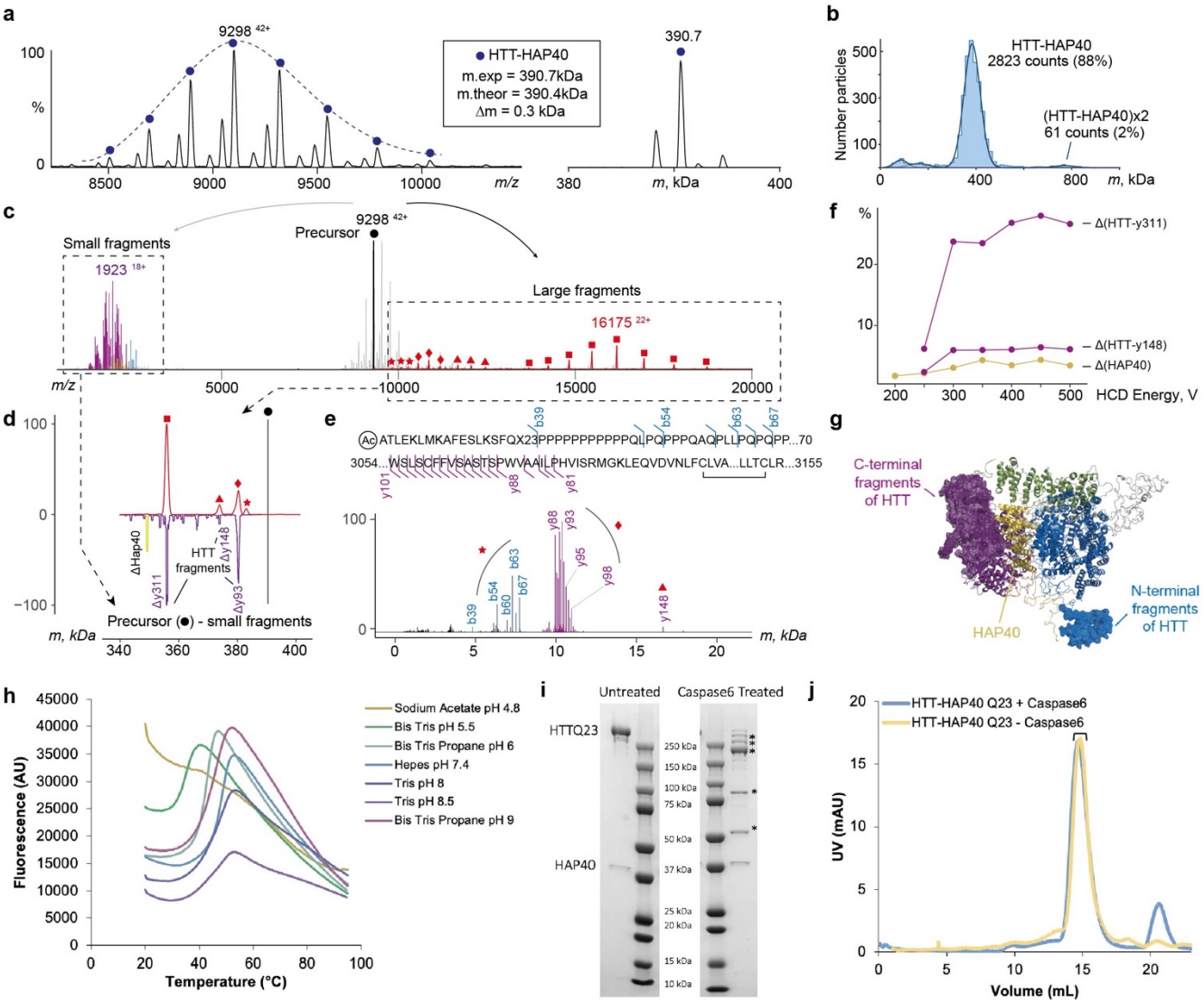


Figure 4. HTT and HAP40 form a very stable non-covalent complex that withstands dissociation.
a Raw native (left) and a deconvoluted zero-charged (right) spectrum of the HTT-HAP40 Q23 complex. **b** Mass profile of HTT-HAP40 complex obtained using mass photometry, showing that the complex is monodisperse. **c** Composite native top-down mass spectrum of the HTT-HAP40 complex demonstrating large (right of the precursor) and small (left of the precursor) dissociation products produced at the highest activation energy. The data reveal that N- and C-terminal fragments of HTT are eliminated from the HTT-HAP40 complex upon collisional activation, whereas the intact HAP40 remains bound. Small fragment peaks are colored following domain colour convention for the HTT-HAP40 complex. **d** Mass distribution of the large HTT-HAP40 fragments, mirrored with the mass distribution of precursor mass subtracted the masses of small fragments. **e** Annotation of small fragments obtained at high-resolution settings and mapping to the sequence of HTT Q23. **f** Energy-resolved plot of fragment abundances: HTT with HAP40 ejected (yellow), HTT upon release of C-terminal fragment y311 or y148 (purple). **g** Structure of HTT-HAP40 complex with eliminated regions highlighted and represented as mesh. Colour-coding is in accordance with the domain colour convention for HTT-HAP40. **h** Assessing HTT-HAP40 Q23 complex stability by measuring transition temperature using DSF in different buffer conditions with 300 mM NaCl. **i** Caspase6 digestion of HTT-HAP40 Q23 proteins assessed by SDS-PAGE and **j** analytical gel filtration. Peak fractions from gel filtration run on SDS-PAGE are indicated.

Polyglutamine expansion modulates the dynamic sampling of conformational space by exon 1

Next, we sought to understand how the disease-causing polyglutamine expansions affect HTT structure. Our structural, biophysical and biochemical data presented so far focus on wildtype HTT (23 glutamines; Q23) and illustrate the importance of HAP40 in stabilising and orienting the HEAT repeat subdomains of HTT. However, 25% of the complex is not resolved in the cryo-EM maps, including many functionally important regions of the protein such as exon 1 (residues 1-90), which harbors the polyglutamine repeat region, and the IDR (residues 407-665). To further investigate the HTT protein structure in its entirety and the influence of polyglutamine expansion within exon 1, we repeated the DSF and proteolysis studies using HTT-HAP40 samples containing either a pathological Huntington's disease HTT with 54 glutamines (Q54), or an HTT with a partially deleted exon 1 (Δexon 1; comprising residues 80-3144, missing N17, polyglutamine and proline-rich domain). We found that neither the Q54 expansion nor the removal of exon 1 had detectable effects on the stability of the HTT-HAP40 complexes compared to the canonical Q23 complex (**Supplementary Figure 3**).

To better describe the structure of exon 1 and the effects of the polyglutamine expansion on the HTT-HAP40 complex, we performed cross-linking mass spectrometry (XL-MS) experiments [36] using the IMAC-enrichable lysine cross-linker, PhoX [37]. For Q23, Q54 and Δexon1 isoforms of HTT-HAP40, we mapped approximately 120 cross-links for each sample (**Supplementary Data File 7**). Importantly, the vast majority of cross-links map to regions unresolved in the cryo-EM maps (**Figure 5a**), thereby providing valuable restraints for structural modeling of a more complete HTT-HAP40 complex. The mean distance of cross-links observed for resolved regions of the cryo-EM model was significantly below the 25 Å distance limit of PhoX in all three datasets (Q23: 7 cross-links – mean distance 13.7 Å; Q54: 11 cross-links – mean distance 14.8 Å; Δexon 1: 12 cross-links – mean distance 14.9 Å; **Supplementary Data File 7**). This, together with mass photometry data of cross-linked HTT-HAP40, indicates that there is a low probability of intermolecular cross-links between HTT molecules, e.g. from aggregation, being included in our datasets (**Supplementary Figure 4a**).

Overall, we obtained very similar cross-link data for the three different HTT-HAP40 constructs (**Figure 5b**). However, of particular note are the large number of exon 1 PhoX cross-links in the HTT-HAP40 Q23 and Q54 samples mediated via lysine-6 or lysine-9 within the N-terminal 17 residues (N17 region) of exon 1. N17 is reported to play key roles for the HTT protein including modulating cellular localisation, aggregation and toxicity [38–40] and is proposed to interact with distal parts of HTT [41].

For both samples (Q23 and Q54), N17 is found to contact several regions of the N-HEAT domain as well as the cryo-EM unresolved N-terminal region of HAP40, via lysine-32 and lysine-40. Interestingly, N17 of Q54 showed additional cross-links to the more distant C-HEAT domain (**Figure 5b**, **Supplementary Figure 4b).** Finally, the largest uninterrupted stretch of the HTT-HAP40 protein which is unresolved in the cryo-EM maps is the IDR. However, only a few PhoX cross-links are detected for it, even though this 258 aa. region harbors 8 lysine residues.

Size-exclusion chromatography multi-angle light scattering (SEC-MALS) analysis of this same series of samples shows no significant difference in mass but does indicate a small shift in the peak for the elution volume of the HTT-HAP40 Δexon 1 complex compared to Q23 and Q54 complex samples (**Figure 6a**). Together with the XL-MS data, this suggests that there are subtle structural differences between the Q23, Q54 and Δexon 1 HTT-HAP40 complexes. To further interpret the cross-linking data

in the context of the 3D structure of the HTT-HAP40 complex, we performed SAXS analysis of our samples to assess any changes to their global structures. We have previously reported SAXS data for HTT-HAP40 Q23 [21]. This revealed that the particle size was significantly larger than the cryo-EM model, which likely accounts for the ~25% of the protein not resolved in cryo-EM maps and therefore not modeled in the structure. Similar analysis of the HTT-HAP40 Q54 and HTT-HAP40 ∆exon 1 and comparison with our previous Q23 data shows that polyglutamine expansion or deletion of exon 1 has only very modest effects on the SAXS profiles (**Figure 6b, c** and **d**). HTT-HAP40 Q54 is slightly larger than the HTT-HAP40 Q23 whereas HTT-HAP40 Δexon 1 samples are slightly smaller, as might be expected, but overall the SAXS determined parameters for the three samples are very similar (**Figure 6e**). In line with that, the SAXS-calculated particle envelopes for the three samples are also very similar in size and shape (**Supplementary Figure 5a**).

Next, we modelled the complete structures of HTT-HAP40, including flexible and disordered regions, integrating our cryo-EM, SAXS and XL-MS data. Coarse-grain modelling molecular dynamics simulations were performed and an ensemble of models that best fit both the cross-linking and SAXS data for HTT–HAP40 was calculated for all three variants of the HTT-HAP40 complex (**Supplementary Figure 5b** and **c**). This modeling approach assumed that the residues with known coordinates in the cryo-EM model form a quasi-rigid complex, whereas the residues with missing coordinates are flexible. As expected from our cross-linking results, the conformations adopted by exon 1 in the ensemble model of Q54 HTT-HAP40 complex are skewed compared to the Q23 ensemble with exon 1 interacting with many more surfaces of the Q54 HTT-HAP40 complex (**Figure 7a**). Mapping our PhoX exon 1 cross-linked residues for each sample to a representative model from each ensemble reveals how exon 1 Q23 cross-links are largely constrained to the N-HEAT domain whereas exon 1 Q54 cross-links are also found on the C-HEAT domain (**Supplementary Figure 4b**). Exon 1 of our HTT-HAP40 Q54 ensemble explores a larger volume of conformational space and this seems to have a knock-on effect on the conformational space occupied by the IDR (**Figure 7b**). Modeling of our HTT-HAP40 structure indicates that the exon 1 region of the Q23 HTT is long enough to make cross-links with the C-HEAT domain, but we do not observe such cross-links in our PhoX datasets (**Supplementary Figure 5d**). This suggests that the additional cross-links observed for the polyglutamine expanded form of HTT-HAP40 may not be driven solely by the length of the exon 1 region. For all ensembles the IDR is differentially constrained and occluded from adopting certain conformations depending on the conformational space occupied by exon 1, suggesting polyglutamine and exon 1-mediated structural changes propagate to the IDR. For the HTT-HAP40 Q54 model ensemble where exon 1 adopts the most diverse conformations, the IDR is the most constrained, occupying a more finite space. However, for the HTT-HAP40 ∆exon 1 model ensemble, the IDR is not occluded and so adopts a much wider range of conformations.

Together, our data suggest that whilst polyglutamine expansion does not affect the core HEAT repeat structure, it does affect the conformational dynamics of not only the exon 1 region but also the IDR.

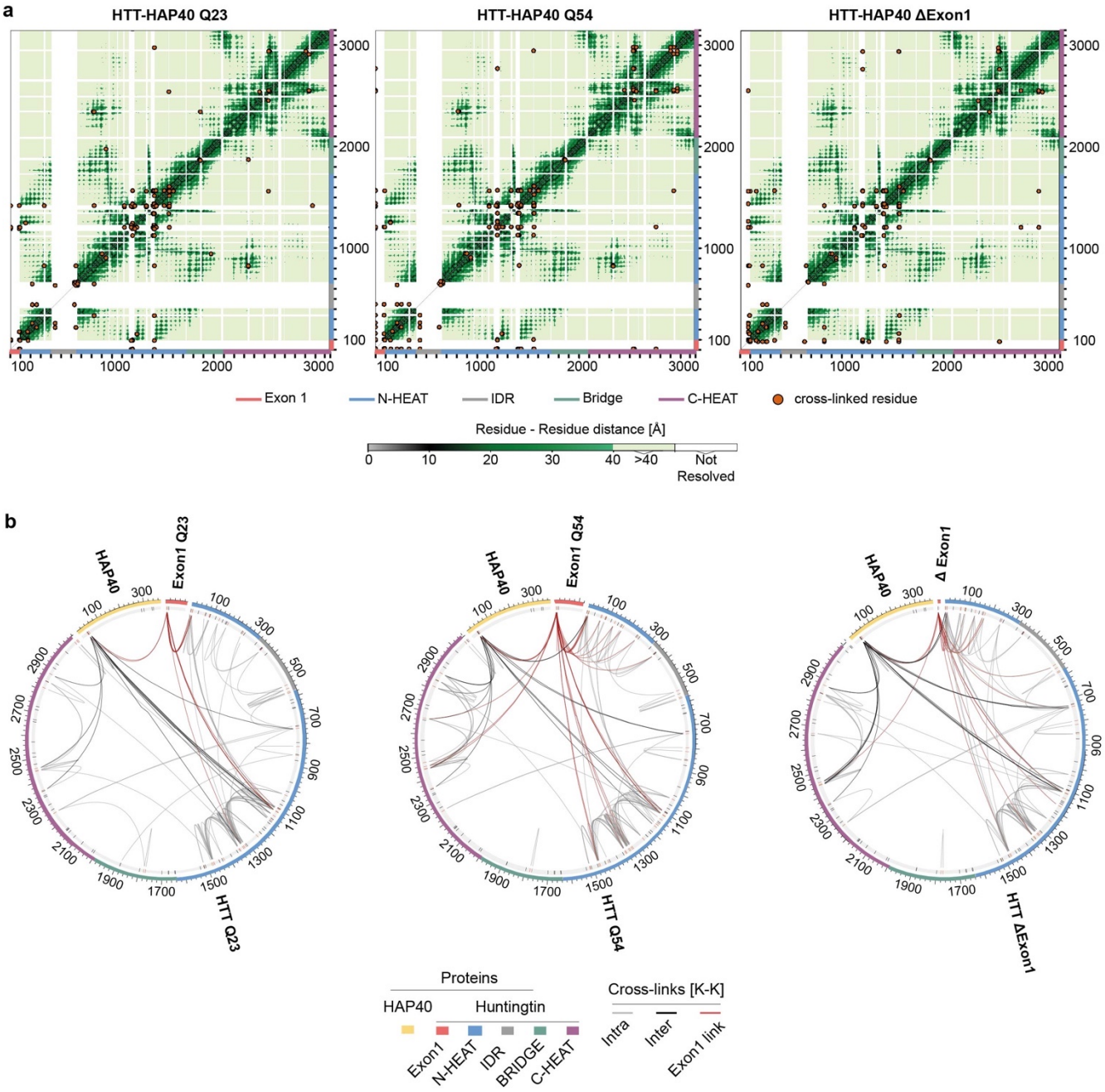
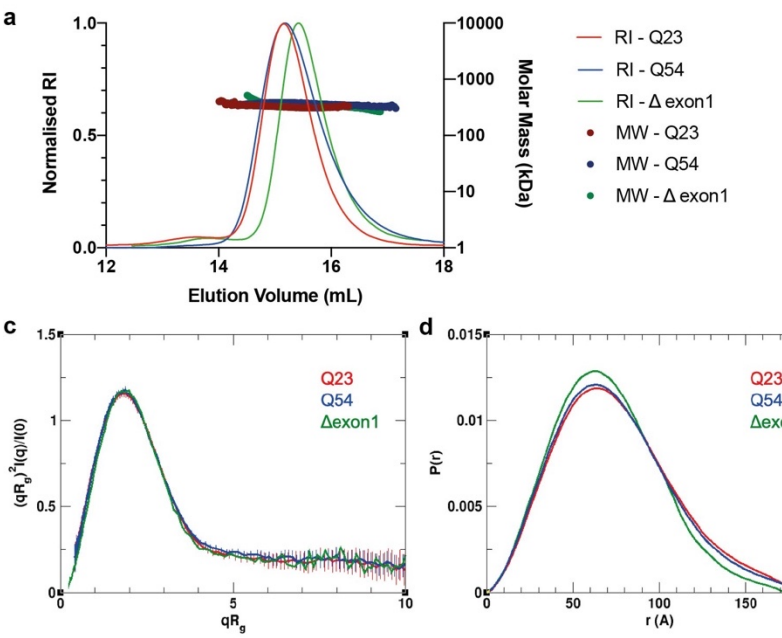


Figure 5. Exon 1 is highly flexible and conformationally dynamic in the context of the full-length protein.

a Mapping cross-linked sites to the HTT-HAP40 sequence of different samples, with cross-linked residue pairs shown as orange circles. Intramolecular distances for HTT-HAP40 (PDBID: 6X9O) shown from grey to green as per the coloured scale bar with unmodelled regions of the protein shown in white. **b** Mapping cross-links to the HTT-HAP40 sequence of different samples, with exon 1 in red, N-HEAT in blue, bridge domain in green, IDR in grey, C-HEAT in purple and HAP40 in yellow. Cross-linked lysine residues are indicated in red and unmodified lysine residues are indicated in black on the numbered sequence. Intermolecular cross-links (HTT-HAP40) are shown in black, intramolecular cross-links (HAP40-HAP40 or HTT-HTT) are shown in grey and exon 1 cross-links are shown in red. All residues following the exon 1 region of the different constructs are numbered the same for clarity.



e

	Rg / Å (Guinier approx.)	Rg / Å (GNOM regularised, smooth fit)	Dmax / Å	MW / kDa predicted (calculated)
Q23	57.7±0.2	57.9±0.4	179	395 (390)
Q54	58.0±0.2	56.8±0.4	182	399 (392)
Δexon1	56.8±7.5	54.1±0.9	176	393 (384)

Figure 6. Polyglutamine expansion or deletion of exon 1 has modest effects on the full-length HTT-HAP40 SAXS profile.

a SEC-MALS analysis of HTT-HAP40 samples Q23 (red), Q54 (blue) and Δexon 1 (green). **b** Experimental SAXS data. **c** Rg-based (dimensionless) Kratky plots of experimental SAXS data for HTT–HAP40 Q23 (red), Q54 (blue) and Δexon 1 (green). **d** Normalized pair distance distribution function P(r) calculated from experimental SAXS data with GNOM for HTT-HAP40 samples. **e** SAXS parameters for data validation and interpretation including radius of gyration (Rg) calculated using Guinier fit in the q range 0.015 < q < 0.025 Å−1, radius of gyration calculated using GNOM, maximum distance between atoms calculated using GNOM, and the molecular mass estimated using SAXSMoW with expected masses from the respective construct sequences shown in the parentheses.

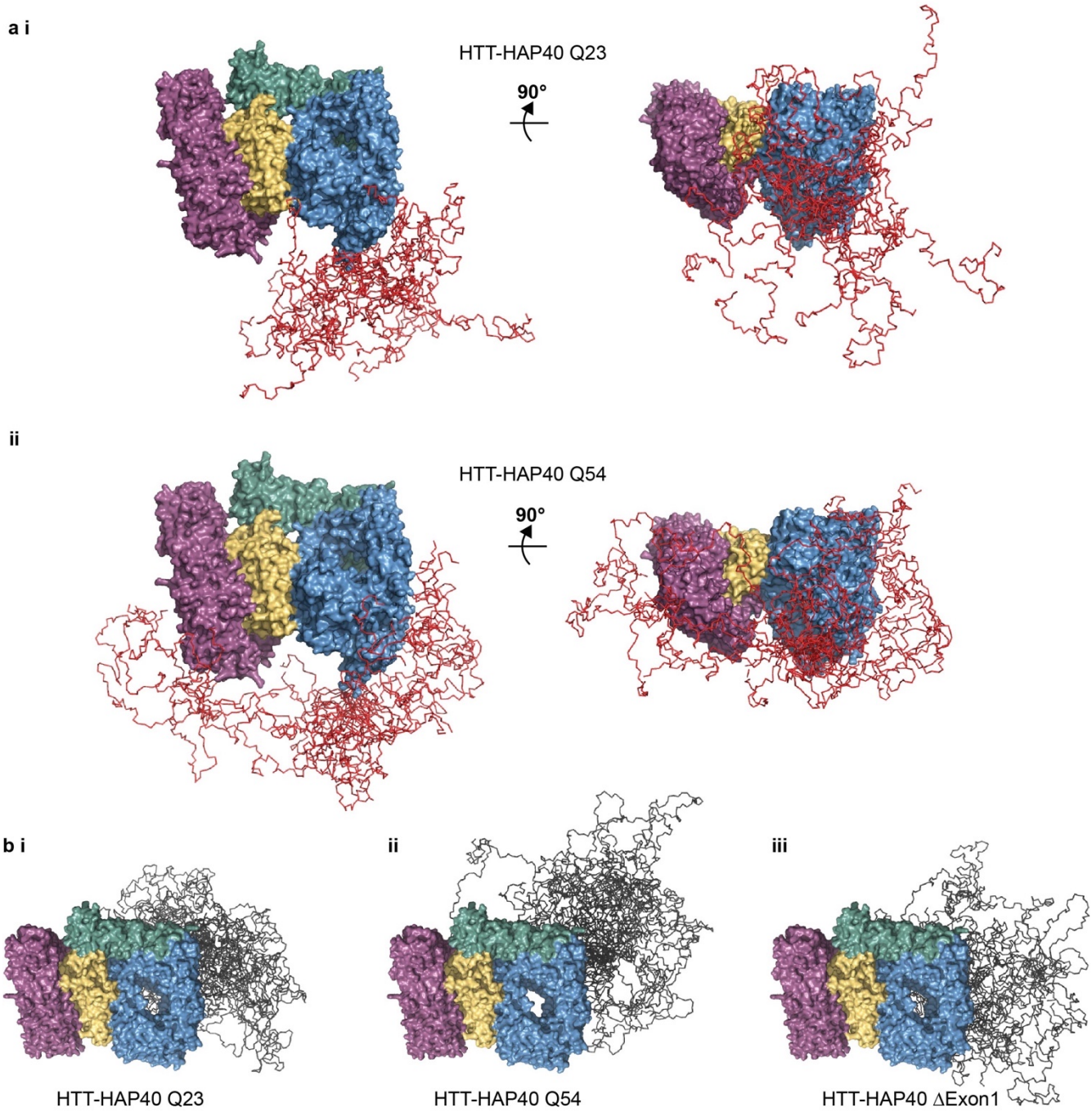


Figure 7. Novel insights from integrated model of full-length HTT-HAP40 combining cryo-EM, SAXS and cross-linking mass spectrometry data.

a Ensemble of models for HTT-HAP40 **i** Q23 and **ii** Q54 showing only the residues defined by the cryo-EM model in surface representation (N-HEAT in blue, bridge domain in green, C-HEAT in purple and HAP40 in yellow) and exon 1 simulated residues in ribbon representation (red). **b** Ensemble of models for HTT-HAP40 **i** Q23, **ii** Q54 and **iii** Δexon 1 showing only the residues defined by the cryo-EM model in surface representation (N-HEAT in blue, bridge domain in green, C-HEAT in purple and HAP40 in yellow) and IDR simulated residues in ribbon representation (grey).

Discussion

We present unprecedented findings for the HTT-HAP40 structure, highlighting the close relationship between HTT and HAP40 as well as unveiling the effect of the polyglutamine expansion, thereby contributing to a richer understanding of HTT and its dependence on HAP40.

HTT is reported to interact with hundreds of different proteins [14] but very few have been validated and the only interaction partner resolved by structural methods is HAP40. HAP40 is thought to have coevolved with HTT [15] and orthologues have been identified in species back to flies [16]. The codependence of HTT and HAP40 is highlighted with our *in vivo* analysis of HTT and HAP40 levels in mice which shows a strong correlation of the two proteins. It remains to be seen if HTT and HAP40 are in fact constitutively bound to each other, or if they may exist independently or in complex with other binding partners. HAP40 plays an important role in stabilising HTT conformation as we have shown with our biophysical and structural comparison of apo and HAP40-bound HTT samples, but the molecular mechanisms of how HAP40 functions in endosome transport [17,18] or modulating HTT toxicity in HD models [16] remains to be determined. Interestingly, despite the exceptional stability of the HTT-HAP40 interaction, complex integrity was not maintained in our DSF assay at low pH, conditions similar to that of the local environment of the endosome. The stabilisation of HTT by HAP40 could be critical for the function of HTT in the stress response to maintain both its structure and function [42].

How polyglutamine expansion of HTT contributes to changes in protein structure-function remains a critical and unanswered question in HD research. Previously, we have observed that changes in polyglutamine tract length seem to have minimal effects on the biophysical properties of HTT and HTT-HAP40 samples [21]. Similarly, in this study, we find no significant differences between our Q23, Q54 and ∆exon 1 HTT-HAP40 samples when assessing monodispersity by mass photometry and native MS; thermal stability in a systematic buffer screen by DSF or stabilisation by proteolysis experiments. The structural differences of Q23, Q54 and ∆exon 1 HTT-HAP40 samples are not resolved within the high-resolution cryo-EM maps we calculated. Our experiments using lower resolution structural methods such as SAXS and mass spectrometry, which do consider the complete protein molecule, also show modest differences between the samples. One way we might rationalise this observation with what we know about HD pathology and huntingtin biology in physiological conditions is that our experimental systems do not capture any subtle, low abundance or slowly occurring differences of the samples which could be important in HD progression that occurs very slowly, over decades of a patient's lifetime. Alternatively, it may be that models of HD pathogenesis which posit that large changes in HTT's globular structure caused by polyglutamine expansion are incorrect.

Notwithstanding the above caveats, our cross-linking mass spectrometry studies provide some of the first insight into the structure of the exon 1 portion of the protein in the context of the full-length, HAP40-bound form of HTT. In both Q23 and Q54 samples, exon 1 appears to be highly dynamic and able to adopt multiple conformations. We demonstrate clear and novel structural differences between the unexpanded and expanded forms of exon 1 in the context of the full-length HTT protein with expanded Q54 forms of exon 1 sampling different conformational space than unexpanded Q23. This is not just due to the additional length of this form of exon 1, conferring a higher degree of flexibility and extension to different regions of the protein but perhaps some biophysical consequence

of a longer polyglutamine tract. This is the opposite of what has been reported for HTT exon 1 protein in isolation, where polyglutamine expansion compacts the exon 1 structure [42–44]. Our data suggest that in the context of the full-length HAP40-bound HTT protein, exon 1 is not compact, but flexible and conformationally dynamic whilst retaining moderate structural organisation. Our modelling studies interestingly suggest that the change in exon 1 conformational sampling upon polyglutamine expansion may have consequent effects on the relative conformations and orientations of the IDR, a novel insight to HTT structure. Both exon-1 and the IDR have been highlighted as functionally important regions of HTT, as sites of dynamic PTMs and protease recognition concentrate in these regions. Our results suggest that structural changes in exon-1 induced by polyglutamine expansion could influence the accessibility of the IDR to partner proteins which modify residues within the IDR, despite the relatively rigid intervening regions between them. The flexibility we observe for exon 1 in both Q23 wildtype and Q54 mutant HTT-HAP40 supports the hypothesis that polyglutamine tracts can function as sensors, sampling and responding to their local environment [45].

Overall, our findings show that HTT is stabilised by interaction with HAP40 through an extensive hydrophobic interface with its distinct HEAT repeat subdomains, creating a highly stable complex. Expanded and unexpanded exon 1 remains highly dynamic in the context of this complex, sampling a vast range of conformational space and interacting with different regions of both HTT and HAP40. We present novel insight into the structural differences of wildtype and mutant HTT, which suggests the conformational constraints of wildtype and mutant exon 1 are significantly different.

Methods

In vivo HTT-HAP40 levels:

Liver tissue was harvested from *Htt*$^{Q111/+}$ mice (JAX:003456) and their WT littermates at 5-6 months of age. To generate samples with genetic reduction of HTT levels in the liver, mice in which the first exon of *Htt* is flanked by LoxP sites [46] were crossed with mice expressing CRE recombinase from the *Alb* promoter (JAX:003574). Liver lysates were prepared for western blotting using non-denaturing lysis buffer (20mM Tris HCl pH8, 127mM NaCl, 1% NP-40, 2mM EDTA), with 50ug of protein separated using 3-8% tris-acetate gels (Invitrogen EA0378) and transferred using an iBlot2 transfer system (Invitrogen IB21001). Probing with antibodies against HTT (Abcam EPR5526; 1:1000) and HAP40 (Novus NBP2-54731; 1:500) was performed with overnight incubation at 4C with gentle shaking, followed by incubation with near infrared secondary antibodies (Licor 926-68073; 1:10,000). Signal was normalized to total protein in the lane (Licor 926-11010). Imaging was performed using a Odyssey imager and signal quantitated using ImageStudio (Licor). All procedures were reviewed and approved by the animal care and use committee at Western Washington University.

Protein expression constructs:

HTT Q23, HTT Q54 and HAP40 constructs used in this study have been previously described [21] and are available through Addgene with accession numbers 111726, 111727 and 124060 respectively. HTT Δexon 1 clones spanning HTT aa. 80-3144 were also cloned into the pBMDEL vector. A PCR product encoding HTT from residues P76 to C3140 was amplified from cDNA (Kazusa clone FHC15881) using primers FWD (ttaagaaggagatatactatgCCGGCTGTGGCTGAGGAGC) and REV (gattggaagtagaggttctctgcGCAGGTGGTGACCTTGTGG). PCR products were inserted using the In-Fusion cloning kit (Clontech) into the pBMDEL that had been linearized with BfuAI. The HTT-coding sequences of expression constructs were confirmed by DNA sequencing. The sequences were also confirmed by Addgene where these reagents have been deposited. This clone is available through Addgene with accession number 162274.

Protein expression and purification:

HTT and HTT-HAP40 protein samples were expressed in insect cells and purified using a similar protocol as previously described [21]. Briefly, Sf9 cells were infected with P3 recombinant baculovirus and grown until viability dropped to 80–85%, normally after ~72 h post-infection. For HTT–HAP40 complex production, a 1:1 ratio of HTT:HAP40 P3 recombinant baculovirus was used for infection. Cells were harvested, lysed with freeze-thaw cycles and then clarified by centrifugation. HTT protein samples were purified by FLAG-affinity chromatography. FLAG eluted samples were bound to Heparin FF cartridge (GE) and washed with 10 CV 20 mM HEPES pH 7.4, 50 mM KCl, 1 mM TCEP, 2.5 % glycerol and eluted with a gradient from 50 mM KCl buffer to 1 M KCl buffer over 10 CV. All samples were purified with a final gel filtration step, using a Superose6 10/300 column in 20 mM HEPES pH 7.4, 300 mM NaCl, 1 mM TCEP, 2.5 % (v/v) glycerol. HTT-HAP40 samples were further purified with an additional Ni-affinity chromatography step prior to gel filtration. Fractions of the peaks corresponding to the HTT monomer or HTT-HAP40 heterodimer were pooled, concentrated, aliquoted and flash frozen prior to use in downstream experiments. Sample purity was assessed by SDS-PAGE. The sample identities were confirmed by native mass spectrometry (**Figure 5**).

SDS-PAGE and western blot analysis

SDS-PAGE and western blot analysis were performed according to standard protocols. Primary antibodies used in western blots are anti-HTT EPR5526 (Abcam), anti-HTT D7F7 (Cell Signaling Technologies) and anti-Flag #F4799 (Sigma). Secondary antibodies used in western blots are goat-anti-rabbit IgG-IR800 (Licor) and donkey anti-mouse IgG-IR680 Licor). Membranes were visualized on an Odyssey® CLx Imaging System (LI-COR).

Differential scanning fluorimetry (DSF) analysis of HTT samples

HTT samples were diluted in different buffer conditions and incubated at room temperature for 15 minutes before the addition of Sypro Orange (Invitrogen) to a final concentration of 5X. The final protein concentration was 0.15 mg/mL. Measurements were performed using a Light Cycler 480 II instrument from Roche Applied Science over the course of 20-95 °C. Temperature scan curves were fitted to a Boltzmann sigmoid function, and the transition temperature values were obtained from the midpoint of the transition.

Caspase6 proteolysis of HTT protein samples

HTT protein samples were mixed with recombinant Caspase6 (Enzo Life Sciences) in a ratio of 100 U caspase6 to 1 pmol of HTT in 20 mM HEPES pH 7.4, 150 mM NaCl, 1 mM TCEP with a final protein concentration of ~1 µM. The reaction and control mixture without caspase6 were incubated at room temperature for 16 hours and then analysed by SDS-PAGE, blue native PAGE and analytical gel filtration using a Superose6 10/300 column in 20 mM HEPES pH 7.4, 150 mM NaCl, 1 mM TCEP.

Cross-linking of HTT-HAP40 samples with PhoX

For cross-linking experiments, HTT-HAP40 samples (HTTQ23-HAP40, HTTQ54-HAP40, HTT Δexon 1-HAP40) were diluted to a protein concentration of 1 mg/1 mL using cross-linking buffer (20 mM Hepes pH 7.4, 300 mM NaCl, 2.5 % glycerol, 1 mM TCEP). HTT-HAP40 samples were treated with an optimised concentration of PhoX cross-linker to avoid protein aggregation (**Supplementary Figure 4a**). After incubation with PhoX (0.5 mM) for 30 min at RT, the reaction was quenched for additional 30 min at RT by the addition of Tris HCl (1 M, pH 7.5) to a final concentration of 50 mM. Protein digestion was performed in 100 mM Tris-HCl, pH 8.5, 1 % SDC, 5 mM TCEP and 30 mM CAA, with the addition of Lys-C and Trypsin proteases (1:25 and 1:100 ratio (w/w)) overnight at 37 °C. The reaction was stopped by addition of TFA to a final concentration of 0.1 % or until pH ~ 2. Next, peptides were desalted using an Oasis HLB plate, before IMAC enrichment of cross-linked peptides like previously described [37].

LC-MS analysis of cross-linked HTT-HAP40 samples

For LC-MS analysis, the samples were re-suspended in 2 % formic acid and analyzed using an UltiMate™ 3000 RSLCnano System (Thermo Fischer Scientific) coupled on-line to either a Q Exactive HF-X (Thermo Fischer Scientific), or an Orbitrap Exploris 480 (Thermo Fischer Scientific). Firstly, peptides were trapped for 5 min in solvent A (0.1 % FA in water), using a 100-µm inner diameter 2-cm trap column (packed in-house with ReproSil-Pur C18-AQ, 3 µm) prior to separation on an analytical column (50 cm of length, 75 µM inner diameter; packed in-house with Poroshell 120 EC-C18, 2.7 µm). Peptides were eluted following a 45 or 55 min gradient from 9-35 % solvent B (80 % ACN, 0.1 % FA), respectively 9- 41 % solvent B. On the Q Exactive HF-X a full scan MS spectra from 375-1600 Da were acquired in the Orbitrap at a resolution of 60,000 with the AGC target set to 3 x 106 and maximum injection time of 120 ms. For measurements on the Orbitrap Exploris 480, a full scan MS spectra from

375-2200 m/z were acquired in the Orbitrap at a resolution of 60,000 with the AGC target set to 2 x 106 and maximum injection time of 25 ms. Only peptides with charged states 3-8 were fragmented, and dynamic exclusion properties were set to n = 1, for a duration of 10 s (Q Exactive HF-X), respectively 15 s (Orbitrap Exploris 480). Fragmentation was performed using in a stepped HCD collision energy mode (27, 30, 33 % Q Exactive HF-X; 20, 28, 36 % Orbitrap Exploris 480) in the ion trap and acquired in the Orbitrap at a resolution of 30,000 after accumulating a target value of 1 x 105 with an isolation window of 1.4 m/z and maximum injection time of 54 ms (Q Exactive HF-X), respectively 55 ms Orbitrap Exploris 480.

Data analysis of HTT-HAP40 cross-links

Raw files for cross-linked HTT-HAP40 samples were analyzed using the XlinkX node [47] in Proteome Discoverer (PD) software suit 2.5 (Thermo Fischer Scientific), with signal to noise threshold set to 1.4. Trypsin was set as a digestion enzyme (max. two allowed missed cleavages), the precursor tolerance set to 10 ppm and the maximum FDR set to 1 %. Additionally, carbamidomethyl modification (Cystein) was set as fixed modification and acetylation (protein N-terminus) and oxidation (Methionine) were set as dynamic modifications. Cross-links obtained for respective HTTQ-HAP40 samples were filtered (only cross-links identified with an XlinkX score > 40 were considered) and further validated using our recently deposited structure of HTTQ23-HAP40 (PDBID: 6X9O) (EMD-22106). Contact maps and circos plots were generated in R (http://www.R-project.org/) using the circlize [48] and XLmaps [49] packages.

Mass photometry

Mass photometry analysis was performed on a Refeyn OneMP instrument (Oxford, UK), which was calibrated using a native marker protein mixture (NativeMark Unstained Protein Standard, Thermo Scientific). The marker contained proteins in the wide mass range up to 1.2 MDa. Four proteins were used to generate a standard calibration curve, with following rounded average masses: 66, 146, 480, and 1048 kDa. The experiments were conducted using glass coverslips, extensively cleaned through several rounds of washing with Milli-Q water and isopropanol. A set of 4-6 gaskets made of clear silicone was placed onto the thoroughly dried glass surface to create wells for sample load. Typically, 1 µL of HTT samples was applied to 19 µL of PBS resulting in a final concentration of ~ 5 nM. Movies consisting of 6000 final frames were recorded using AcquireMP software at a 100 Hz framerate. Particle landing events were automatically detected amounting to ~ 3000 per acquisition. The data was analyzed using DiscoverMP software. Average masses of HTT proteins and HTT-HAP40 complexes were determined by taking the value at the mode of the normal distribution fitted into the histograms of particle masses. Finally, probability density function was calculated and drawn over the histogram to produce the final mass profile. Measurement and analysis of mass photometry data were done for the following samples: HTT-Q23-HAP40, HTT-Q54-HAP40, and HTT-Δexon 1-HAP40.

Intact mass and middle-down MS sample preparation

Sample preparation: Samples containing HTT-HAP40 complexes were digested using human Caspase6 (Enzo Life Sciences, Farmingdale, USA) by adding 200 U of the enzyme to the 20 µg of the protein. The mixture was stored in PBS for 24 hours. Following the digestion, samples were diluted to the final concentration of 500 ng/µL with 2 % formic acid. Approximately 2 µg of the sample were injected for a single intact mass LC-MS or middle-down LC-MS/MS experiment.

LC-MS(/MS) for intact and middle-down MS

Produced peptides of HTT were separated using a Vanquish Flex UHPLC (Thermo Fisher Scientific, Bremen, Germany) coupled on-line to an Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Fisher Scientific, San Jose, USA) via reversed-phase analytical column (MAbPac, 1 mm × 150 mm, Thermo Fisher Scientific). The column compartment and preheater were kept at 80°C during the measurements to ensure efficient unfolding and separation of the analyzed peptides. Analytes were separated and measured for 22 min at a flow rate of 150 µL/min. Elution was conducted using A (Milli-Q H2O/0.1 % CH_2O_2) and B (C_2H_3N/0.1 % CH_2O_2) mobile phases. In the first minute B was increased from 10 to 30 %, followed by 30 to 57% B gradient over 14 minutes, 1 min 57 to 95 % B ramp-up, 95 % B for 1 min, and equilibration of the column at 10 % B for 4 min.

During data acquisition, Lumos Fusion instrument was set to Intact Protein and Low Pressure mode. MS1 resolution of 7,500 (determined at 200 m/z and equivalent to 16 ms transient signal length) was used, which enables optimal detection of protein ions above 30 kDa in mass. Mass range of 500-3,000 m/z, the automatic gain control (AGC) target of 250 %, and a max injection time (IT) of 50 ms were used for recording of MS1 scans. 2 µscans were averaged in the time domain and recorded for the 7,500 resolution scans during the LC-MS experiment and 5 µscans for when tandem MS (MS/MS) was performed. MS/MS scans were recorded at a resolution setting of 120,000 (determined at 200 m/z and equivalent to 16 ms transient signal length), 10,000 % AGC target, 250 ms max IT, and five µscans, for the single most abundant peak detected in the preceding MS1 scan. The selected ions were mass-isolated by a quadrupole in a 4 m/z window and accumulated to an estimate of 5e6 ions prior to the gas-phase activation. Two separate LC-MS/MS runs were recorded per sample with either higher-energy collisional dissociation (HCD) or electron transfer dissociation (ETD) used for fragmentation. For ETD following parameters were used: ETD reaction time – 16 ms, max IT of the ETD reagent – 200 ms, and the AGC target of the ETD reagent – 1e6. For HCD, 30 V activation energy was used. MS/MS scans were acquired with the minimum intensity of the precursor set to 5e4 and the range of 350-5000 m/z using quadrupole in the high mass isolation mode.

Data analysis of intact and middle-down MS

LC-MS data were deconvoluted with ReSpect algorithm in BioPharma Finder 3.2 (Thermo Fisher Scientific, San Jose, USA). ReSpect parameters: precursor m/z tolerance – 0.2 Th, target mass – 50 kDa, relative abundance threshold – 0 %, mass range – 3-100 kDa; tolerance – 30 ppm, charge range – 3-100. MS1 and MS2 masses were recalibrated using an external calibrant mixture of intact proteins (PiercePierce™ Intact Protein Standard Mix, Thermo Scientific) measured before and after each HTT sample. Iterative sequence adjustments of putative HTT peptides was done until the exact precursor and fragment masses matched to determine a final set of HTT peptides generated by Caspase6 enzyme. HCD fragments of HTT peptides were used solely to confirm identified sequences. Phosphorylation was matched as 80 Da variable modification mass, added to the mass of the identified HTT peptides. Visualization was done in R extended with ggplot2 package.

Native (top-down) MS sample preparation

Samples were stored at -80°C in the buffer containing 20 mM HEPES pH 7.4, 300 mM NaCl, 2.5 % (v/v) glycerol, 1 mM TCEP. Approximately 40 µg of the HTT-Q23, HTT-Q54, HTT-Δexon 1, and their respective complexes with Hap40 protein were buffer-exchanged into 150 mM aqueous ammonium acetate (pH=7.5) by using P-6 Bio-Spin gel filtration columns (Bio-rad, Veenendaal, the Netherlands).

The protein's resulting concentration was estimated to be ~2-5 µM before native MS analysis. For the recording of denaturing MS, samples were spiked with formic acid to the final concentration of 2% right before the MS measurement.

Native (top-down) data acquisition

HTT-containing samples were directly injected into a Q Exactive Ultra-High Mass Range (UHMR) Orbitrap mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) using in-house pulled and gold-coated borosilicate capillaries. Following mass spectrometer parameters were used: capillary voltage – 1.5 kV, positive ion mode, source temperature – 250 °C, S-lens RF level – 200, injection time – mostly 200 ms, noise level parameter – 3.64. In-source trapping with a desolvation voltage of -100 V was used to desolvate the proteinaceous ions efficiently. No additional acceleration voltage was used in the back-end of the instrument. The automatic gain control (AGC) was switched to fixed. Resolutions of 4,375 and 8,750 (both at m/z = 200 Th) were used, representing 16 and 32 ms transient, respectively. Ion guide optics and voltage gradient throughout the instrument were manually adjusted for optimal transmission and detection of HTT and HTT-HAP40 ions. The higher-energy collisional dissociation (HCD) cell was filled with Nitrogen, and the trapping gas pressure was set to 3 or 4 setting value, corresponding to ~2e-10 – 4e-10 mBar for the ultra-high vacuum (UHV) readout of the instrument. The instrument was calibrated in the *m/z* range of interest using a concentrated aqueous cesium iodide (CsI) solution. Acquisition of the spectra was usually performed by averaging 100-200 µscans in the time domain. Peaks corresponding to the protein complex of interest were isolated with a 20 Th window for single charge state isolation and a 2000 Th window for charge-state ensemble isolation. In both cases, isolated HTT-HAP40 ions were investigated for dissociation using elevated HCD voltages, with direct eV setting varied in the range 1-500 V. For detection of high-m/z dissociation product ions, mass analyzer detection mode and transmission RF settings were set to “high m/z”. For detection of low-m/z fragment ions, all relevant instrument settings were set to “low m/z”, and the instrument resolution was increased to 140,000 (at m/z = 200 Th).

Data Analysis for native (top-down) MS

Raw native MS and high-m/z native top-down MS data were processed with UniDec [50] to obtain zero-charged mass spectra. Native top-down MS data recorded with high resolution (140,000) were deconvoluted using the Xtract algorithm within FreeStyle software (1.7SP1; Thermo Fisher Scientific). The resulting zero-charge fragments were matched to the theoretical fragments produced for HTT and Hap40 using in-house scripts with 5 ppm mass tolerance. Final visualization was performed in R extended with ggplot2 library.

Cryo-EM sample preparation and data acquisition

HTT was diluted to 0.4 mg/ml in 20 mM HEPES pH 7.5, 300 mM NaCl, 1 mM TCEP and adsorbed to glow-discharged holey carbon-coated grids (Quantifoil 300 mesh, Au R1.2/1.3) for 10 s. Grids were then blotted with filter paper for 2 s at 100 % humidity at 4 °C and frozen in liquid ethane using a Vitrobot Mark IV (Thermo Fisher Scientific).

HTT-HAP40 was diluted to 0.2 mg/ml in 25 mM HEPES pH 7.4, 300 mM NaCl, 0.025 % w/v CHAPS, 1 mM DTT and adsorbed onto gently glow-discharged suspended monolayer graphene grids

(Graphenea) for 60 s. Grids were then blotted with filter paper for 1 s at 100 % humidity, 4 °C and frozen in liquid ethane using a Vitrobot Mark IV (Thermo Fisher Scientific).

Data were collected in counting mode on a Titan Krios G3 (FEI) operating at 300 kV with a BioQuantum imaging filter (Gatan) and K2 direct detection camera (Gatan) at 165,000x magnification, pixel size of 0.822 Å. Movies were collected over 32 fractions at a dose rate of 6.0 e-/Å^2/s, exposure time of 8 s, resulting in a total dose of 48.0 e-/Å^2.

Cryo-EM data processing

For apo HTT, patched motion correction and dose weighting were performed using MotionCor implemented in RELION 3.0 [51]. Contrast transfer function parameters were estimated using CTFFIND4 [52]. Particles were picked in SIMPLE 3.0 [53] and processed in RELION 3.0. 669 movies were collected in total and 108,883 particles extracted. Particles were subjected to one round of reference-free 2D classification against 100 classes (k = 100) using a soft circular mask of 180 Å in diameter in RELION. A subset of 25,424 particles were recovered at this stage and subjected to 3D auto-refinement in RELION using a 40 Å lowpass-filtered map of HTT-HAP40 (EMDB 3984) as initial reference. This generated a ~12 Å map based on gold-standard Fourier shell correlation curves using the 0.143 criterion as calculated within RELION.

For HTT-HAP40 (**Supplementary Figure 1**), 15,003 movies were processed in real time using the SIMPLE 3.0 pipeline, using SIMPLE-unblur for patched motion correction, SIMPLE-CTFFIND for patched CTF estimation and SIMPLE-picker for particle picking. After initial 2D classification in SIMPLE 3.0 using the cluster2D_stream module (k =500), cleaned particles were imported into RELION and subjected to reference-free 2D classification (k = 200) using a 180 Å soft circular mask. An *ab initio* map, generated from a selected subset of particles (372,226), was subsequently lowpass filtered to 40 Å and used as reference for coarse-sampled (7.5°) 3D classification (k = 4) with a 180 Å soft spherical mask against the same particle subset. Particles (102,729) belonging to the most defined, highest resolution class were selected for 3D auto-refinement against its corresponding map, lowpass filtered to 40 Å, using a soft mask covering the protein which generated a 3.5 Å volume. This map was lowpass filtered to 40 Å and used as initial reference for a multi-step 3D classification (k = 5, 15 iterations at 7.5° followed by 5 iterations at 3.75°), with 180 Å soft spherical mask, against the full cleaned dataset of 2,240,373 particles. Selected particles (647,468) from the highest resolution class were subjected to masked 3D auto-refinement against its reference map, lowpass filtered to 15 Å, yielding a 3.1 Å volume. CTF refinement using per-particle defocus plus beamtilt estimation further improved map quality to 3.0 Å. Bayesian particle polishing followed by an additional round of CTF refinement with per-particle defocus plus beamtilt estimation on a larger box size (448 x 448) generated a final volume with global resolution of 2.6 Å as assessed by Gold standard Fourier shell correlations using the 0.143 criterion within RELION. Map local resolution estimation was calculated within Relion (**Supplementary Figure 1**). Additional rounds of 3D classification using either global/local searches or classification only without alignment did not improve map quality.

Model building and refinement

The model for HTT-HAP40 (**Supplementary Table 1**) was generated by rigid body fitting the 4 Å HTT-HAP40 model [20] (PDBID: 6EZ8) into our globally-sharpened, local resolution filtered 2.6 Å map followed by multiple rounds of manual real-space refinement using Coot v. 0.95 [54] and automated

real-space refinement in PHENIX v. 1.18.2-38746 [55] using secondary structure, rotamer and Ramachandran restraints. HTT-HAP40 model was validated using MolProbity [56] within PHENIX. Figures were prepared using UCSF ChimeraX v.1.1 [57] and PyMOL v.2.4.0 (The PyMOL Molecular Graphics System, v.2.0; Schrödinger).

SAXS data collection and analysis

SAXS experiments were performed at beamline 12-ID-B of the Advanced Photon Source (APS) at Argonne National Laboratory. The energy of the X-ray beam was 13.3 keV (wavelength λ = 0.9322 Å), and two setups (small- and wide-angle X-ray scattering) were used simultaneously to cover scattering q ranges of $0.006 < q < 2.6$ Å^{-1}, where $q = (4\pi/\lambda)\sin\theta$, and 2θ is the scattering angle. For HTT-HAP40 Q54, thirty two-dimensional images were recorded for buffer or sample solutions using a flow cell, with an exposure time of 0.8 s to reduce radiation damage and obtain good statistics. The flow cell is made of a cylindrical quartz capillary 1.5 mm in diameter and 10 µm wall thickness. Concentration-series measurements for this sample were carried out at 300 K with concentrations of 0.5, 1.0, and 2.0 mg/ml, in 20 mM HEPES, pH 7.5, 300 mM NaCl, 2.5% (v/v) glycerol, 1 mM TCEP. No radiation damage was observed as confirmed by the absence of systematic signal changes in sequentially collected X-ray scattering images. The 2D images were corrected for solid angle of each pixel, and reduced to 1D scattering profiles using the Matlab software package at the beamlines. The 1D SAXS profiles were grouped by sample and averaged.

For HTT-HAP40 Δexon 1, data were collected using an in-line FPLC AKTA micro setup with a Superose6 Increase 10/300 GL size exclusion column in 20 mm HEPES, pH 7.5, 300 mm NaCl, 2.5% (v/v) glycerol, 1 mm TCEP. A 150uL sample loop was used and the stock sample concentration was 5 mg/ml. The sample passed through the FPLC column and was fed to the flow cell for SAXS measurements. The SAXS data were collected every 2 seconds and the X-ray exposure time was set to 0.75 seconds. Only the SAXS data collected above the half maximum of the elution peak, about 50-100 frames, were averaged and for further analysis. Background data were collected before and after the peak (each 100 frames), while data before the peak were found better and used for the background subtraction.

SAXS data were analyzed with the software package ATSAS 2.8 . The experimental radius of gyration, R_g, was calculated from data at low q values using the Guinier approximation. The pair distance distribution function, P(r), the maximum dimension of the protein, D_{max}, and R_g in real space were calculated with the indirect Fourier transform using the program GNOM [60]. Estimation of the molecular weight of samples was obtained by both SAXMOW [61,62] and by using volume of correlation, Vc [63]. The theoretical scattering intensity of the atomic structure model was calculated using FoXS [64]. *Ab-initio* shape reconstructions (molecular envelopes) were performed using both bead modeling with DAMMIF [65] and calculating 3D particle electron densities directly from SAXS data with DENSS [66].

Coarse-grained molecular dynamics simulations

We used a Gō-like coarse-grained model of HTT/HAP40 for structural modeling of the complex as it was described previously [21]. We build two different models that are based on two experimental EM structures of the complex (PDBIDs: 6EZ8 and 6X9O, respectively). We used experimentally observed cross-links to improve the sampling of the flexible regions of the model by introducing in the force field a distance restraint term given by the following potential:

$$V_{XL}(t) = \sum_{k=1}^{N_{XL}} \delta^{k}_{\xi(t)} V_l^k \; ; \qquad V_l^k = K_{XL}/(1 + e^{-\beta(l_k(t) - l_0)})$$

The sum is over all cross-links, N_{XL} *i*s the number of cross-links; l_k is the C_α-C_α distance for residues involved in *kth* cross-link; l_0 = 25 is the upper bound for PhoX cross-links; β = 0.5 is the slope of the sigmoidal function; K_{XL} = 10 kcal/mol is the force constant; δ_i^k is the Kronecker delta; and $\xi(t)$ is the random digital number selected from the interval [1, N_{XL}]. We chose to keep active only about $N_{XL}/3$ randomly selected restraints, numbers $\xi(t)$, that are updated every τ_{XL} = 0.5 ns during the MD simulation.

The goodness-of-fit of an ensemble of structural models of the complex to the SAXS data was evaluated by comparing an ensemble average profile, $I_{avrg}(q)$, with the experimental one. $I_{avrg}(q)$ was calculated either by performing simple averaging of model's theoretical scattering intensities over MD trajectory or by selecting optimal ensemble using SES method [67]. Theoretical scattering profiles for each conformation in the MD trajectory were calculated in the *q* range $0 < q < 0.30$ Å^{-1} using FoXS [64].

Size-exclusion chromatography multi angle light scattering (SEC-MALS)

The absolute molar masses and mass distributions of purified protein samples of HTT-HAP40 Q23, HTT-HAP40 Q54 and HTT-HAP40 Δexon 1 at 1 mg/ml were determined using SEC-MALS. Samples were injected through a Superose 6 10/300 GL column (GE Healthcare) equilibrated in 20 mm HEPES, pH 7.5, 300 mm NaCl, 2.5% (v/v) glycerol, 1 mm TCEP followed in-line by a Dawn Heleos-II light scattering detector (Wyatt Technologies) and a 2414 refractive index detector (Waters). Molecular mass calculations were performed using ASTRA 6.1.1.17 (Wyatt Technologies) assuming a dn/dc value of 0.185 ml/g.

In silico analysis of the HTT-HAP40 protein complex structure

HTT-HAP40 models were analysed using Pymol [68] and APBS [69]. For conservation analysis, HTT and HAP40 orthologues were extracted from Ensembl, parsed to remove low quality or partial sequences and then aligned using Clustal [70]. Multiple sequence alignments were then analysed using Consurf [71] and conservation scores mapped to the HTT-HAP40 (PDBID: 6X9O) structure in Pymol. Ligandable pocket analysis was completed as previously reported [72]. Briefly, HTT-HAP40 model pdb files were loaded in ICM (Molsoft, San Diego). Proteins were protonated, optimal positions of added polar hydrogens were generated, correct orientation of side-chain amide groups for glutamine and asparagine and most favourable histidine isomers were identified. The PocketFinder algorithm implemented in ICM, which uses a transformation of the Lennard-Jones potential to identify ligand binding envelopes regardless of the presence of bound ligands, was then applied [73]. Residues with side-chain heavy atoms within 2.8Å of the molecular envelope were identified as lining the pocket.

Acknowledgements

We acknowledge the use of the SAXS Core Facility of the Center for Cancer Research (CCR), NCI, National Institutes of Health. NCI SAXS Core is funded by FNLCR contract HHSN261200800001E and the intramural research program of the NIH, NCI, CCR. This research used 12-ID-B beamline of the Advanced Photon Source, a United States Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357.

This research was supported by the CHDI Foundation (RJH, CHA, JBC), the Huntington Society of Canada (RJH, CHA), the Wellcome Trust #219477 (SML, JD) and the EU Horizon 2020 program INFRAIA project Epic-XS Project 823839 (JFH, ST, AJRH). RJH is the recipient of the Huntington's Disease Society of America Berman Topper Career Development Fellowship.

The Structural Genomics Consortium is a registered charity (no: 1097737) that receives funds from AbbVie, Bayer AG, Boehringer Ingelheim, Genentech, Genome Canada through Ontario Genomics Institute [OGI-196], the EU and EFPIA through the Innovative Medicines Initiative 2 Joint Undertaking [EUbOPEN grant 875510], Janssen, Merck KGaA (aka EMD in Canada and US), Pfizer, Takeda and the Wellcome Trust [106169/ZZ14/Z].

Author Contributions

RJH conceived the project, designed and conducted experiments, analysed and interpreted data, supervised the project and wrote the manuscript. JD, JFH, ST, AL, JPC, MS and XZ designed and conducted experiments, analysed and interpreted data and contributed to drafting and editing the manuscript. MMS, AH, AS and PL conducted experiments and analysed data. AJRH, JBC, CHA. SML and LF supervised the work, analysed and interpreted data and contributed to drafting and editing the manuscript.

The authors declare no competing interests.

Materials and correspondence

All expression constructs are available through Addgene.
Cryo-EM maps can be downloaded at EMDB 22106 and model coordinates at PDBID 6X9O.

All correspondence and requests for materials should be sent to RJH (Rachel.Harding@utoronto.ca) or CHA (Cheryl.Arrowsmith@uhnresearch.ca).

References

1. Donaldson, J., Powell, S., Rickards, N., Holmans, P. & Jones, L. What is the Pathogenic CAG Expansion Length in Huntington's Disease? *J. Huntingt. Dis.* **10**, 175–202 (2021).
2. Saudou, F. & Humbert, S. The Biology of Huntingtin. *Neuron* **89**, 910–926 (2016).
3. Harding, R. J. & Tong, Y. Proteostasis in Huntington's disease: disease mechanisms and therapeutic opportunities. *Acta Pharmacol. Sin.* **39**, 754–769 (2018).
4. Koyuncu, S., Fatima, A., Gutierrez-Garcia, R. & Vilchez, D. Proteostasis of Huntingtin in Health and Disease. *Int. J. Mol. Sci.* **18**, (2017).
5. Gao, R. *et al.* Mutant huntingtin impairs PNKP and ATXN3, disrupting DNA repair and transcription. *eLife* **8**, e42988 (2019).
6. Poplawski, G. H. D. *et al.* Injured adult neurons regress to an embryonic transcriptional growth state. *Nature* **581**, 77–82 (2020).
7. Carmo, C., Naia, L., Lopes, C. & Rego, A. C. Mitochondrial Dysfunction in Huntington's Disease. *Adv. Exp. Med. Biol.* **1049**, 59–83 (2018).
8. Vitet, H., Brandt, V. & Saudou, F. Traffic signaling: new functions of huntingtin and axonal transport in neurological disease. *Curr. Opin. Neurobiol.* **63**, 122–130 (2020).
9. Smith-Dijak, A. I., Sepers, M. D. & Raymond, L. A. Alterations in synaptic function and plasticity in Huntington disease. *J. Neurochem.* **150**, 346–365 (2019).
10. McColgan, P. & Tabrizi, S. J. Huntington's disease: a clinical review. *Eur. J. Neurol.* **25**, 24–34 (2018).
11. Maiuri, T. *et al.* Huntingtin is a scaffolding protein in the ATM oxidative DNA damage response complex. *Hum. Mol. Genet.* **26**, 395–406 (2017).
12. Rui, Y.-N. *et al.* Huntingtin functions as a scaffold for selective macroautophagy. *Nat. Cell Biol.* **17**, 262–275 (2015).
13. Shirasaki, D. I. *et al.* Network organization of the huntingtin proteomic interactome in mammalian brain. *Neuron* **75**, 41–57 (2012).
14. Wanker, E. E., Ast, A., Schindler, F., Trepte, P. & Schnoegl, S. The pathobiology of perturbed mutant huntingtin protein-protein interactions in Huntington's disease. *J. Neurochem.* **151**, 507–519 (2019).
15. Seefelder, M. *et al.* The evolution of the huntingtin-associated protein 40 (HAP40) in conjunction with huntingtin. *BMC Evol. Biol.* **20**, 162 (2020).
16. Xu, S. *et al.* HAP40 is a conserved central regulator of Huntingtin and a specific modulator of mutant Huntingtin toxicity. *bioRxiv* 2020.05.27.119552 (2020) doi:10.1101/2020.05.27.119552.
17. Pal, A., Severin, F., Lommer, B., Shevchenko, A. & Zerial, M. Huntingtin-HAP40 complex is a novel Rab5 effector that regulates early endosome motility and is up-regulated in Huntington's disease. *J. Cell Biol.* **172**, 605–618 (2006).
18. Pal, A., Severin, F., Höpfner, S. & Zerial, M. Regulation of endosome dynamics by Rab5 and Huntingtin-HAP40 effector complex in physiological versus pathological conditions. *Methods Enzymol.* **438**, 239–257 (2008).
19. Peters, M. F. & Ross, C. A. Isolation of a 40-kDa Huntingtin-associated Protein. *J. Biol. Chem.* **276**, 3188–3194 (2001).
20. Guo, Q. *et al.* The cryo-electron microscopy structure of huntingtin. *Nature* (2018) doi:10.1038/nature25502.

21. Harding, R. J. *et al.* Design and characterization of mutant and wild-type huntingtin proteins produced from a toolkit of scalable eukaryotic expression systems. *J. Biol. Chem.* jbc.RA118.007204 (2019) doi:10.1074/jbc.RA118.007204.
22. Jung, T. *et al.* The Polyglutamine Expansion at the N-Terminal of Huntingtin Protein Modulates the Dynamic Configuration and Phosphorylation of the C-Terminal HEAT Domain. *Structure* **28**, 1035-1050.e8 (2020).
23. Vijayvargia, R. *et al.* Huntingtin's spherical solenoid structure enables polyglutamine tract-dependent modulation of its structure and function. *eLife* **5**, e11184 (2016).
24. Boatz, J. C. *et al.* Protofilament Structure and Supramolecular Polymorphism of Aggregated Mutant Huntingtin Exon 1. *J. Mol. Biol.* **432**, 4722–4744 (2020).
25. Falk, A. S. *et al.* Structural Model of the Proline-Rich Domain of Huntingtin Exon-1 Fibrils. *Biophys. J.* **119**, 2019–2028 (2020).
26. Matlahov, I. & van der Wel, P. C. Conformational studies of pathogenic expanded polyglutamine protein deposits from Huntington's disease. *Exp. Biol. Med. Maywood NJ* **244**, 1584–1595 (2019).
27. Ratovitski, T. *et al.* Post-Translational Modifications (PTMs), Identified on Endogenous Huntingtin, Cluster within Proteolytic Domains between HEAT Repeats. *J. Proteome Res.* (2017) doi:10.1021/acs.jproteome.6b00991.
28. Schilling, B. *et al.* Huntingtin Phosphorylation Sites Mapped by Mass Spectrometry MODULATION OF CLEAVAGE AND TOXICITY. *J. Biol. Chem.* **281**, 23686–23697 (2006).
29. Tabrizi, S. J., Ghosh, R. & Leavitt, B. R. Huntingtin Lowering Strategies for Disease Modification in Huntington's Disease. *Neuron* **101**, 801–819 (2019).
30. Wheeler, V. C. *et al.* Length-Dependent Gametic CAG Repeat Instability in the Huntington's Disease Knock-in Mouse. *Hum. Mol. Genet.* **8**, 115–122 (1999).
31. Evers, M. M. *et al.* Making (anti-) sense out of huntingtin levels in Huntington disease. *Mol. Neurodegener.* **10**, 21 (2015).
32. Bogdanos, D. P., Gao, B. & Gershwin, M. E. Liver Immunology. in *Comprehensive Physiology* 567–598 (American Cancer Society, 2013). doi:10.1002/cphy.c120011.
33. Huang, B. *et al.* PolyQ expansion does not alter the Huntingtin-HAP40 complex. *bioRxiv* 2021.02.02.429316 (2021) doi:10.1101/2021.02.02.429316.
34. Huang, B. *et al.* Scalable Production in Human Cells and Biochemical Characterization of Full-Length Normal and Mutant Huntingtin. *PLOS ONE* **10**, e0121055 (2015).
35. Graham, R. K. *et al.* Cleavage at the Caspase-6 Site Is Required for Neuronal Dysfunction and Degeneration Due to Mutant Huntingtin. *Cell* **125**, 1179–1191 (2006).
36. Liu, F., Lössl, P., Scheltema, R., Viner, R. & Heck, A. J. R. Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification. *Nat. Commun.* **8**, 15473 (2017).
37. Steigenberger, B., Pieters, R. J., Heck, A. J. R. & Scheltema, R. A. PhoX: An IMAC-Enrichable Cross-Linking Reagent. *ACS Cent. Sci.* **5**, 1514–1522 (2019).
38. Gu, X. *et al.* N17 Modifies mutant Huntingtin nuclear pathogenesis and severity of disease in HD BAC transgenic mice. *Neuron* **85**, 726–741 (2015).
39. Jayaraman, M. *et al.* Kinetically competing huntingtin aggregation pathways control amyloid polymorphism and properties. *Biochemistry* **51**, 2706–2716 (2012).
40. Maiuri, T., Woloshansky, T., Xia, J. & Truant, R. The huntingtin N17 domain is a multifunctional CRM1 and Ran-dependent nuclear and cilial export signal. *Hum. Mol. Genet.* **22**, 1383–1394 (2013).
41. Caron, N. S., Desmond, C. R., Xia, J. & Truant, R. Polyglutamine domain flexibility mediates the proximity between flanking sequences in huntingtin. *Proc. Natl. Acad. Sci.* **110**, 14610–14615 (2013).

42. Nath, S., Munsie, L. N. & Truant, R. A huntingtin-mediated fast stress response halting endosomal trafficking is defective in Huntington's disease. *Hum. Mol. Genet.* **24**, 450–462 (2015).
43. Bravo-Arredondo, J. M. *et al.* The folding equilibrium of huntingtin exon 1 monomer depends on its polyglutamine tract. *J. Biol. Chem.* **293**, 19613–19623 (2018).
44. Newcombe, E. A. *et al.* Tadpole-like Conformations of Huntingtin Exon 1 Are Characterized by Conformational Heterogeneity that Persists regardless of Polyglutamine Length. *J. Mol. Biol.* **430**, 1442–1458 (2018).
45. Warner, J. B. *et al.* Monomeric Huntingtin Exon 1 Has Similar Overall Structural Features for Wild-Type and Pathological Polyglutamine Lengths. *J. Am. Chem. Soc.* **139**, 14456–14469 (2017).
46. Gerbich, T. M. & Gladfelter, A. S. Moving beyond disease to function: Physiological roles for polyglutamine-rich sequences in cell decisions. *Curr. Opin. Cell Biol.* **69**, 120–126 (2021).
47. Dragatsis, I., Levine, M. S. & Zeitlin, S. Inactivation of Hdh in the brain and testis results in progressive neurodegeneration and sterility in mice. *Nat. Genet.* **26**, 300–306 (2000).
48. Klykov, O. *et al.* Efficient and robust proteome-wide approaches for cross-linking mass spectrometry. *Nat. Protoc.* **13**, 2964–2990 (2018).
49. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
50. Schweppe, D. K., Chavez, J. D. & Bruce, J. E. XLmap: an R package to visualize and score protein structure models based on sites of protein cross-linking. *Bioinformatics* **32**, 306–308 (2016).
51. Marty, M. T. *et al.* Bayesian Deconvolution of Mass and Ion Mobility Spectra: From Binary Interactions to Polydisperse Ensembles. *Anal. Chem.* **87**, 4370–4376 (2015).
52. Zivanov, J., Nakane, T. & Scheres, S. H. W. A Bayesian approach to beam-induced motion correction in cryo-EM single-particle analysis. *IUCrJ* **6**, 5–17 (2019).
53. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
54. Caesar, J. *et al.* SIMPLE 3.0. Stream single-particle cryo-EM analysis in real time. *J. Struct. Biol. X* **4**, 100040 (2020).
55. Brown, A. *et al.* Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Crystallogr. D Biol. Crystallogr.* **71**, 136–153 (2015).
56. Afonine, P. V. *et al.* Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. Sect. Struct. Biol.* **74**, 531–544 (2018).
57. Prisant, M. G., Williams, C. J., Chen, V. B., Richardson, J. S. & Richardson, D. C. New tools in MolProbity validation: CaBLAM for CryoEM backbone, UnDowser to rethink "waters," and NGL Viewer to recapture online 3D graphics. *Protein Sci.* **29**, 315–329 (2020).
58. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).
59. Franke, D. *et al.* ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J. Appl. Crystallogr.* **50**, 1212–1225 (2017).
60. Svergun, D., Barberato, C. & Koch, M. H. J. CRYSOL – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *J. Appl. Crystallogr.* **28**, 768–773 (1995).
61. Fischer, H. *et al.* Determination of the molecular weight of proteins in solution from a single small-angle X-ray scattering measurement on a relative scale. *J. Appl. Crystallogr.* **43**, 101–109 (2010).
62. Piiadov, V., Ares de Araújo, E., Oliveira Neto, M., Craievich, A. F. & Polikarpov, I. SAXSMoW 2.0: Online calculator of the molecular weight of proteins in dilute solution from experimental SAXS data measured on a relative scale. *Protein Sci. Publ. Protein Soc.* **28**, 454–463 (2019).

963 63. Rambo, R. P. & Tainer, J. A. Accurate assessment of mass, models and resolution by small-
964 angle scattering. *Nature* **496**, 477–481 (2013).
965 64. Schneidman-Duhovny, D., Hammel, M. & Sali, A. FoXS: a web server for rapid computation and
966 fitting of SAXS profiles. *Nucleic Acids Res.* **38**, W540–W544 (2010).
967 65. Franke, D. & Svergun, D. I. DAMMIF, a program for rapid ab-initio shape determination in
968 small-angle scattering. *J. Appl. Crystallogr.* **42**, 342–346 (2009).
969 66. Grant, T. D. Ab initio electron density determination directly from solution scattering data.
970 *Nat. Methods* **15**, 191–193 (2018).
971 67. Berlin, K. *et al.* Recovering a Representative Conformational Ensemble from Underdetermined
972 Macromolecular Structural Data. *J. Am. Chem. Soc.* **135**, 16595–16609 (2013).
973 68. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. (2015).
974 69. Dolinsky, T. J. *et al.* PDB2PQR: expanding and upgrading automated preparation of
975 biomolecular structures for molecular simulations. *Nucleic Acids Res.* **35**, W522–W525 (2007).
976 70. Madeira, F. *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids*
977 *Res.* **47**, W636–W641 (2019).
978 71. Ashkenazy, H. *et al.* ConSurf 2016: an improved methodology to estimate and visualize
979 evolutionary conservation in macromolecules. *Nucleic Acids Res.* **44**, W344–W350 (2016).
980 72. Yazdani, S. *et al.* The SARS-CoV-2 replication-transcription complex is a priority target for
981 broad-spectrum pan-coronavirus drugs. *bioRxiv* 2021.03.23.436637 (2021)
982 doi:10.1101/2021.03.23.436637.
983 73. An, J., Totrov, M. & Abagyan, R. Pocketome via comprehensive identification and classification
984 of ligand binding envelopes. *Mol. Cell. Proteomics MCP* **4**, 752–761 (2005).
985