5. Figures

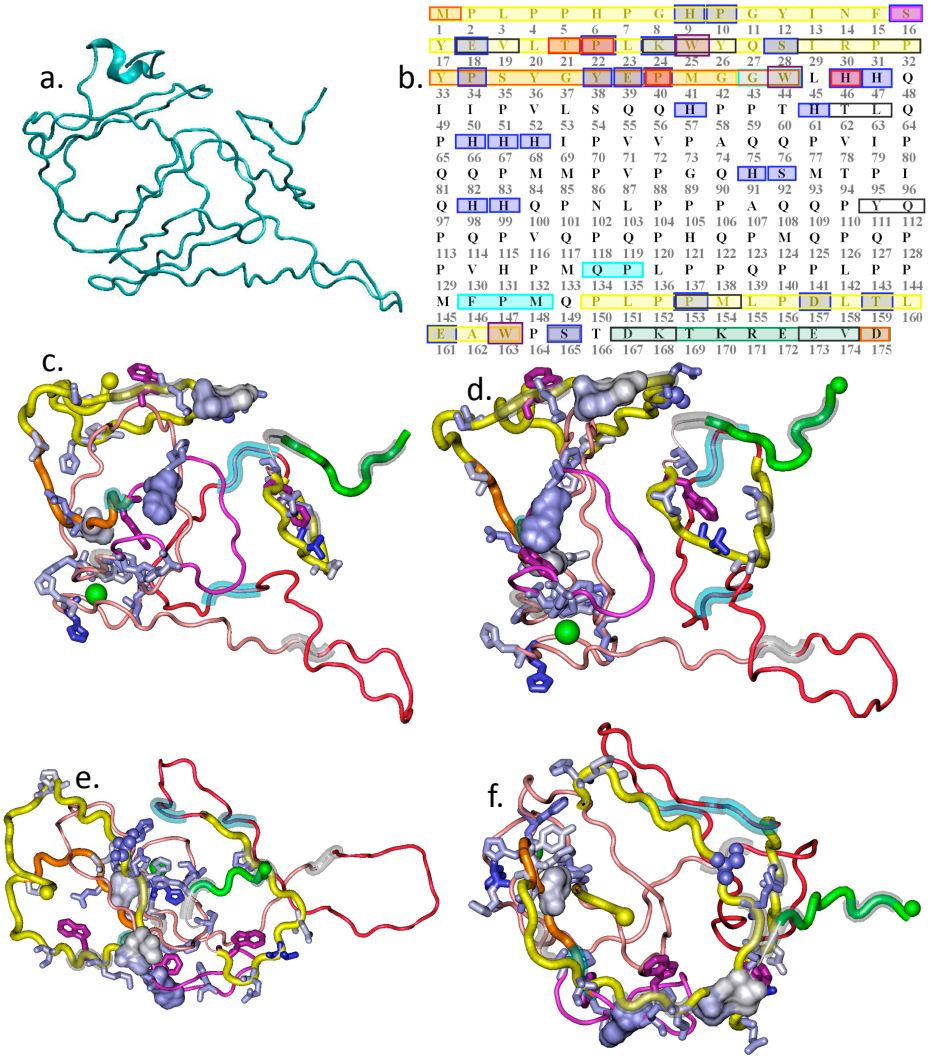

Figure 1. Atomic level resolution of AMELX structure. a. An atomic model for the structure of AMELX H175. **b.** AMELX sequence map of structural and functional data used as criteria to select the AMELX structure from many models, concatenated from the literature (detailed in table 1). Various perspectives of the selected AMELX model are shown with selection criteria mapped onto the structure. **c.** perspective to maximize separation of A and B domains (main chain shown as yellow tube), carboxy terminal (green tube), and HΦ regions (proximal HΦ shown as purple tube, central HΦ as pink tube, distal HΦ as red tube). **d.** five degree offset from c. **e,f.** roughly orthagonal perspectives from c. Residues predicted for highest functional importance (side chains shown and colored from highest predicted function as deep blue to gray for 15th percentile). The top calcium ion binding site prediction is shown as a green sphere. Non-specific protease cleavage sites are shown as transparent cyan tubes (Moradian-Oldak et al., 2001). MMP-20 cleavage sites are shown as transparent black tubes (Ryu et al., 1999). Each side chain atom of the phosphoserine side chain is shown as a small blue sphere (Fincham et al., 1994). The ATMP region is shown as an orange tube (Beniash et al., 2005; Moradian-Oldak et al., 2002; Paine and Snead, 1997). Three freely rotatable tryptophans side chains are shown in magenta (Oobatake et al., 2006). The surface of side chains are shown for amelogenesis imperfecta-related missense mutation sites with darker blue indicating higher predicted functional importance (Collier et al., 1997; Hart et al., 2002; Kida et al., 2007; Lench and Winter, 1995).

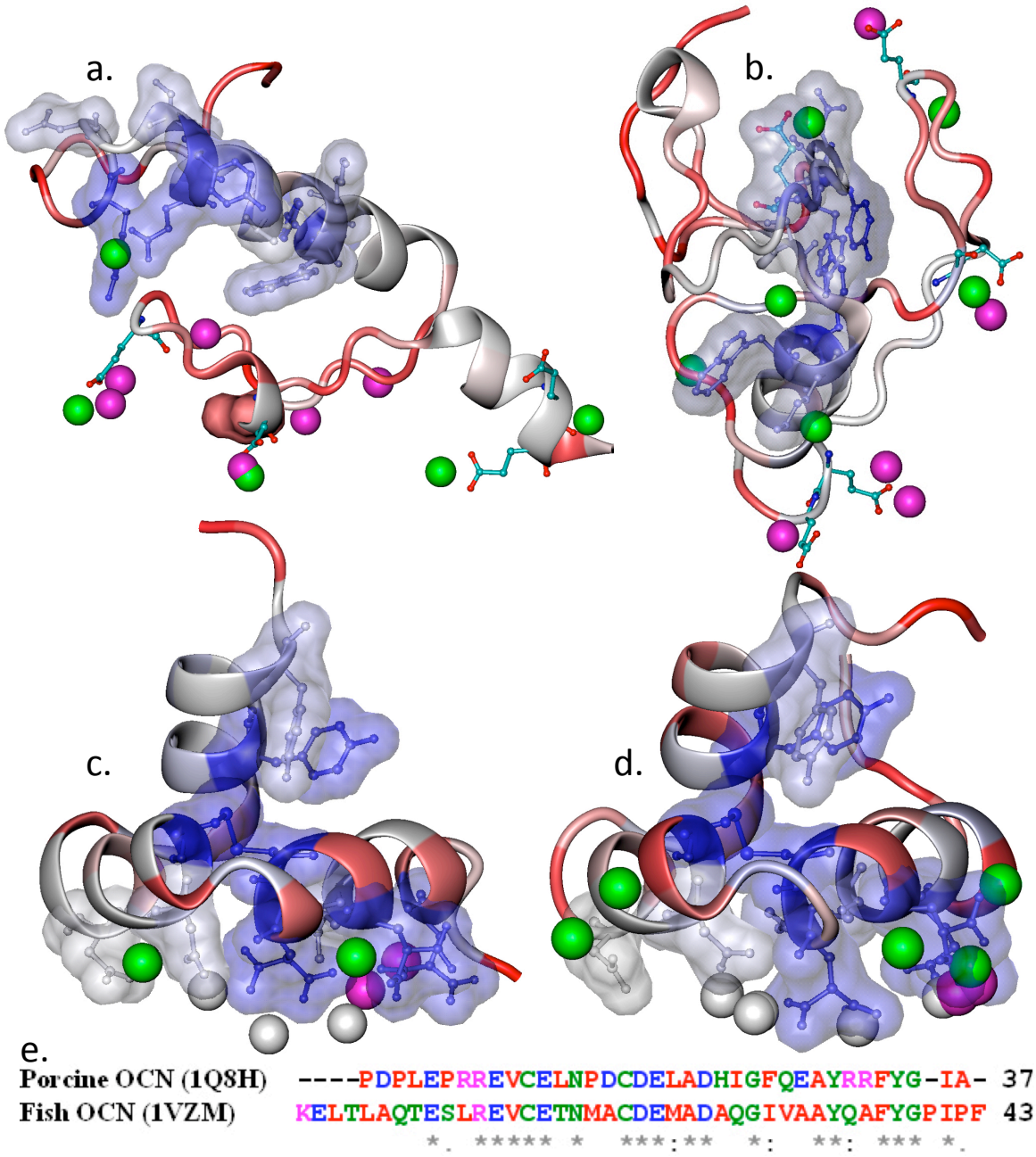

Figure 2. LRAP and osteocalcin structure, function, and interactions. Main chains are colored by predicted functional importance for each residue, with blue for highest function and red for lowest, and side chains are shown for the top 15%. Predicted metal ion binding sites are shown as spheres, with calcium in green, and magnesium in purple. Predicted structures are shown for **a.** LRAP-4 and **b.** LRAP+4. Crystal structures of **c.** porcine (PDB identifier 1q8h; calcium) and **d.** fish osteocalcin (PDB identifier 1vzm; magnesium) were used as positive controls for specific placement of externally bound metal ions (white spheres) thought to represent the hydroxyapatite binding mechanism. **e.** ClustalW alignment for the sequences of the proteins in c and d demonstrates 51% sequence identity.

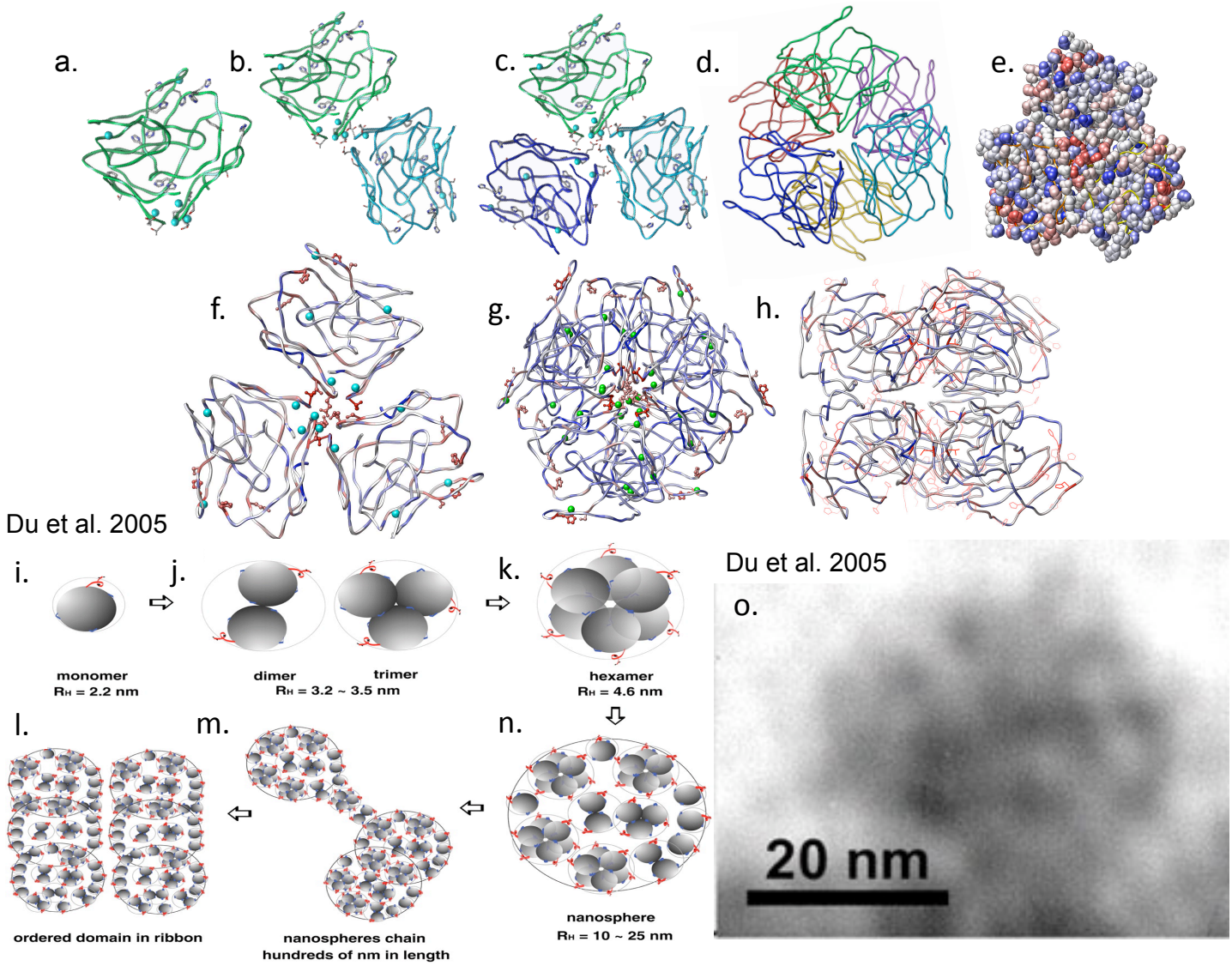

Figure 3. AMELX oligomer assembly. We built an AMELX oligomer construct by docking monomer forms according to the established protein self interaction regions. I tucked in extraneous loops and applied molecular dynamics energy minimization after each movement. Panels a-c and f-h show side chains in red for the top 15% MFS scores. Panels a-c and f-g show calcium ion binding prediction sites (green spheres). **a.** The AMELX monomer (diameter ~5nm). **b.** A model of the dimer begs incorporation of a third monomer. **c.** The trimers can be joined to sequester hydrophobicity. **d.** Each main chain in the hexamer is colored differently to clarify the stacked offset, cartoonized in k. **e.** The hexamer surface shown with hydrophobic surface patches in red and hydrophilic patches in blue. Absence of large surface patches suggest relative stability for this form in solution. **f.** A split hexamer after calcium ion docking shows a central functional site with high MFS score residues that interacts favorably with calcium. **g.** Axial and **h.** side views of the hexamer. **i-o.** Figures taken from Du *et al.* (2005). Each grey sphere depicts **i.** a monomer, **j.** joining to form dimer and trimer, **k.** which then join to form a hexamer, **n.** which assembles into nanospheres, and **m,l.** nanosphere chains thereafter. The AMELX oligomerization models concur. **o.** a transmission electron microscope image of an AMELX nanosphere with discernable semi-globular constituents.

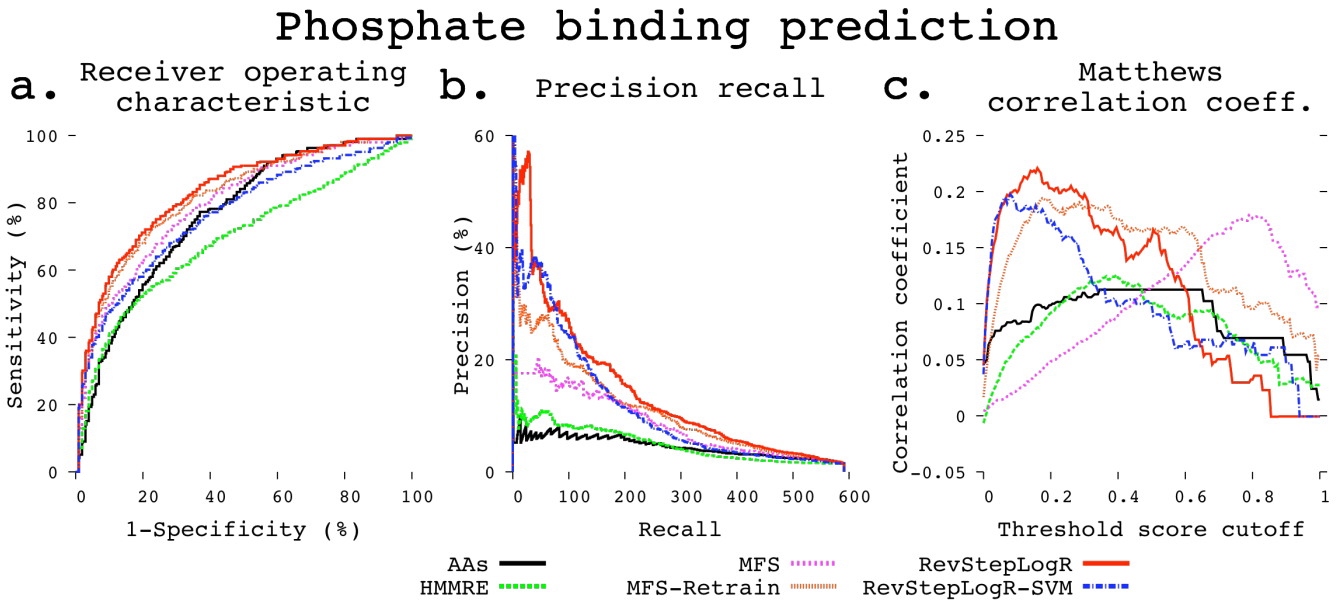

Figure 4. Prediction of phosphate binding residues. a. ROC, **b.** PR, **c.** MCC distribution. Prediction accuracy is improved by new methods to predict function from sequence inferred features of protein structure (RevStepLogR; red) over simply retraining the original MFS parameters for phosphate binders (MFS-Retrain; orange): sequence conservation (HMMRE; green), evolutionary conservation (not shown), and amino acid type (AAs; black). The RevStepLogR precision recall curve shows 57% precision for the top 49 predictions, a success for this "needle in a haystack problem."

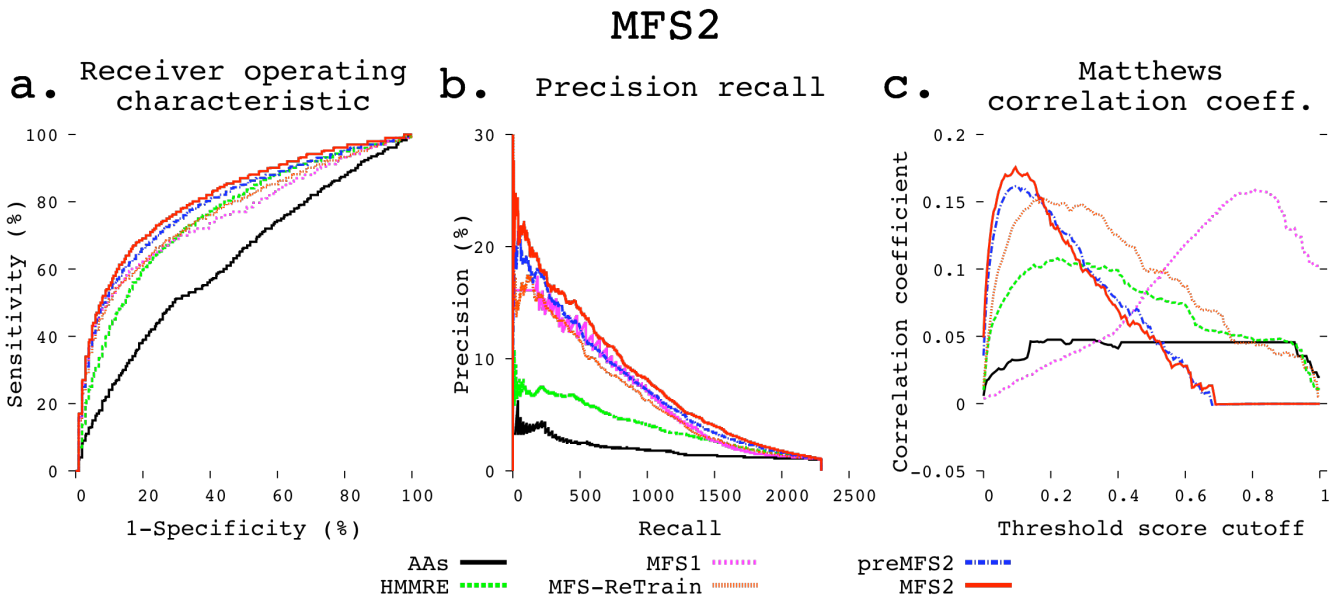

Figure 5. Improvement in general function prediction. Analysis displayed as in figure 4. Heuristic inferences of structure (MFS2, red) improve functional residue prediction compared to sequence conservation (HMMRE, green), the original MFS (purple; Wang et al., 2008), retraining with the same parameters (MFS-ReTrain; orange), and limiting methods to those described in chapter 2 (preMFS2, blue).

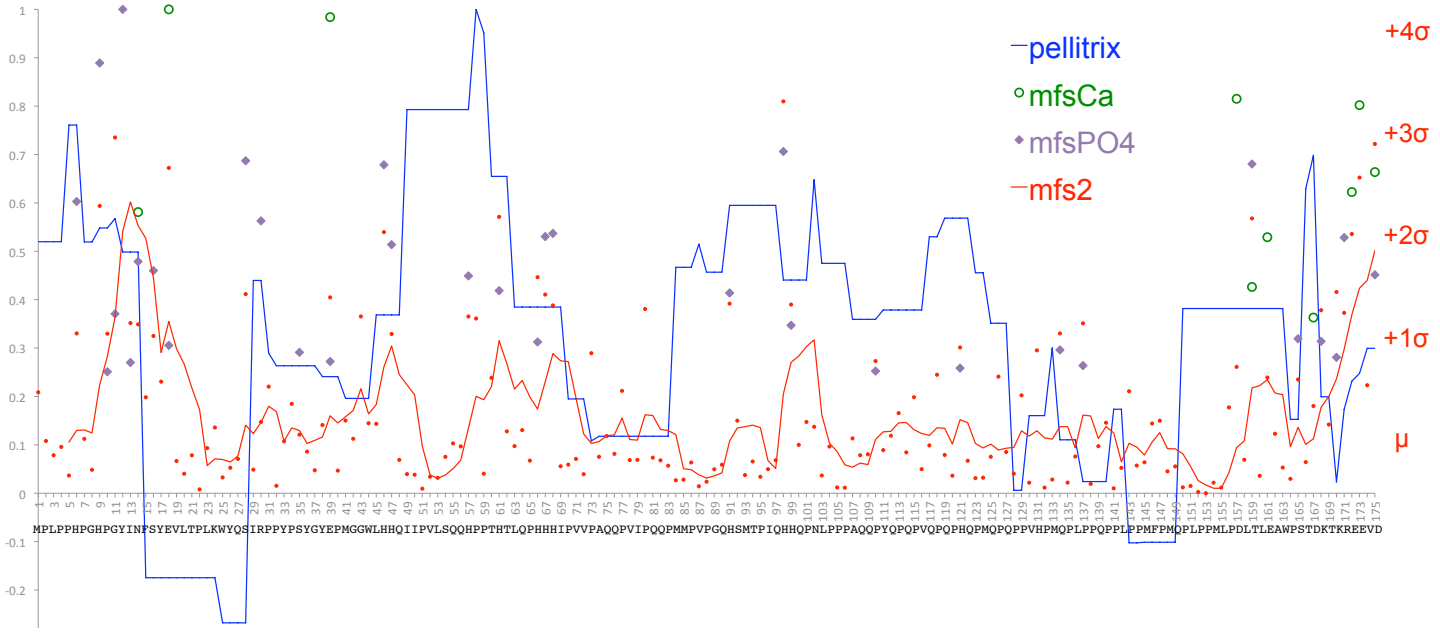

Figure 6. AMELX meta-functional signatures. Individual mfs2 scores of AMELX H175 are shown as red dots, with the five residue average shown as a red line to group functional regions. Supra-threshold mfsCa scores shown as green circles and mfsPO4 as purple diamonds. Enamel pellicle peptide similarity (pellitrix) scores shown as blue line. Mean (μ) and standard deviation (σ) for mfs2 scores demarcated at right. Superimposition of predicted functions or lack thereof models the type of function carried out by the region and residue. For example high predicted general function (mfs2), clusters of phosphate binders (mfsPO4), and similarity to pellicle peptides suggests interaction with mature or maturing hydroxyapatite for residues 46-68 and 91-99; high function (mfs2), phosphate binders (mfsPO4), and calcium binders (mfsCa) but dissimilarity to pellicle peptides suggests function nucleation for 12-18; while low function, no binders, and dissimilarity to pellicle peptides describes the known specific proteolytic cleavage site 146-148.

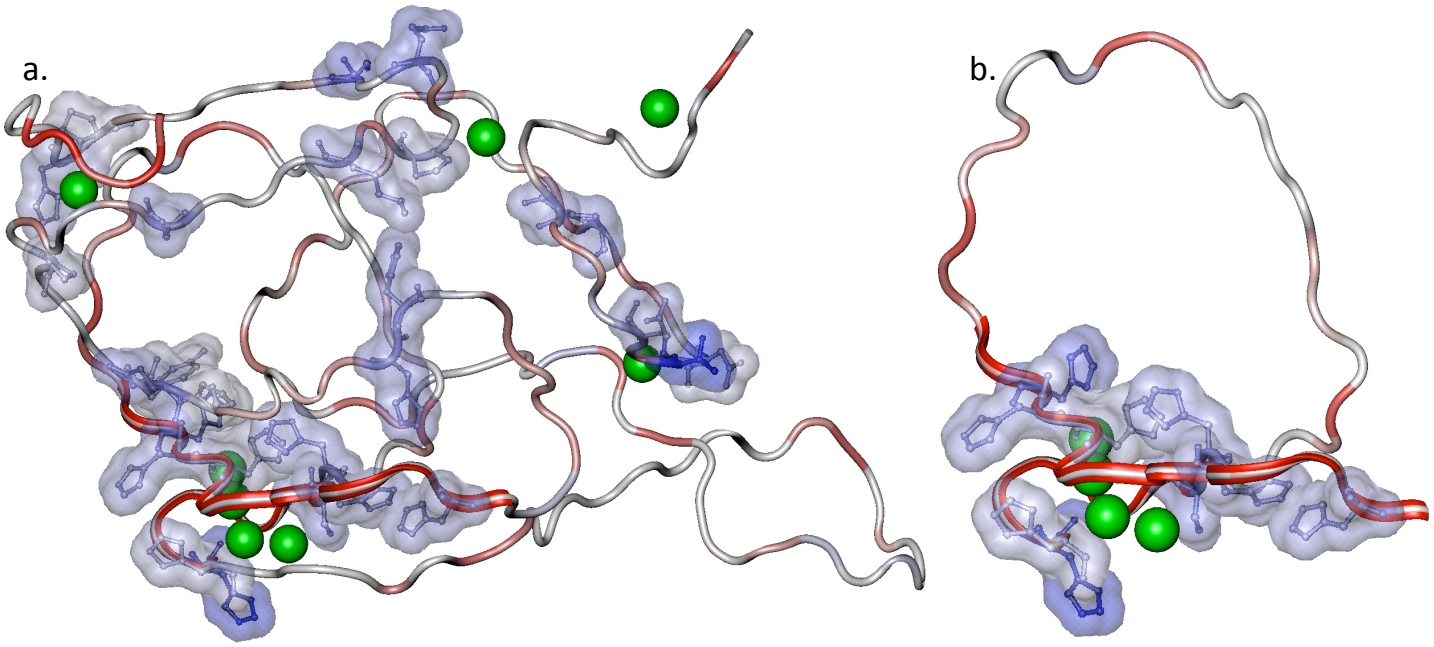

Figure 7. Spatial concordance for functional importance and structure-based calcium ion binding predictions. a. Predicted functional importance of each residue in the AMELX H175 selected structure model is represented by main chain color, with dark blue indicating most important to function and dark red the least. Fourteen of the twenty eight most important residues cluster together (side chains shown for these top 15% residues), and circumscribe four of eight calcium ion binding prediction sites (green spheres). These four calcium ions occur along a linear arrangement with one at each orifice of a pore-like structure. **b.** The consensus region appears to be separable as a functional peptide.

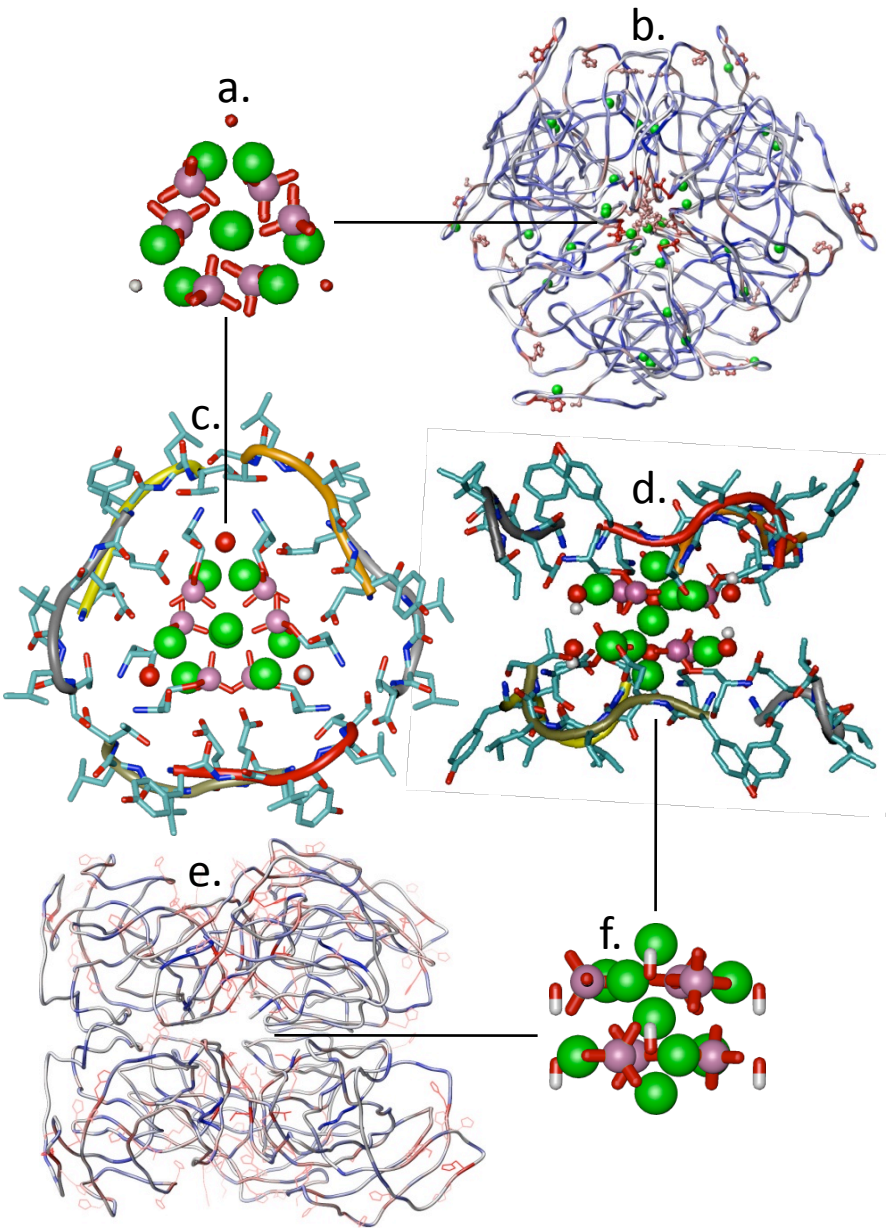

Figure 8. Similar symmetry between the AMELX hexamer and the hydroxyapatite unit cell. **a.** Axial and **f.** side views of the hydroxyapatite unit cell. **b.** Axial and **e.** side views of the AMELX hexamer. **b.** The many calcium binding sites found by SOAK in the middle of AMELX suggests a significant function. The size of the interior cavity would not allow interaction with any substantial portion of the hydroxyapatite surface. However, **b,e.** the symmetry of the hexamer model mimics **a,f.** that of the hydroxyapatite unit cell. Thus it is possible that this emergent functional site brings together calcium, phosphate, and hydroxyl ions to form new seed crystals. Calcium atoms in green, phosphates in purple, oxygen in red, hydrogen in white. **c.** Axial and **d.** side views of the YINLSYE region rebuilt around the hydroxyapatite unit cell.

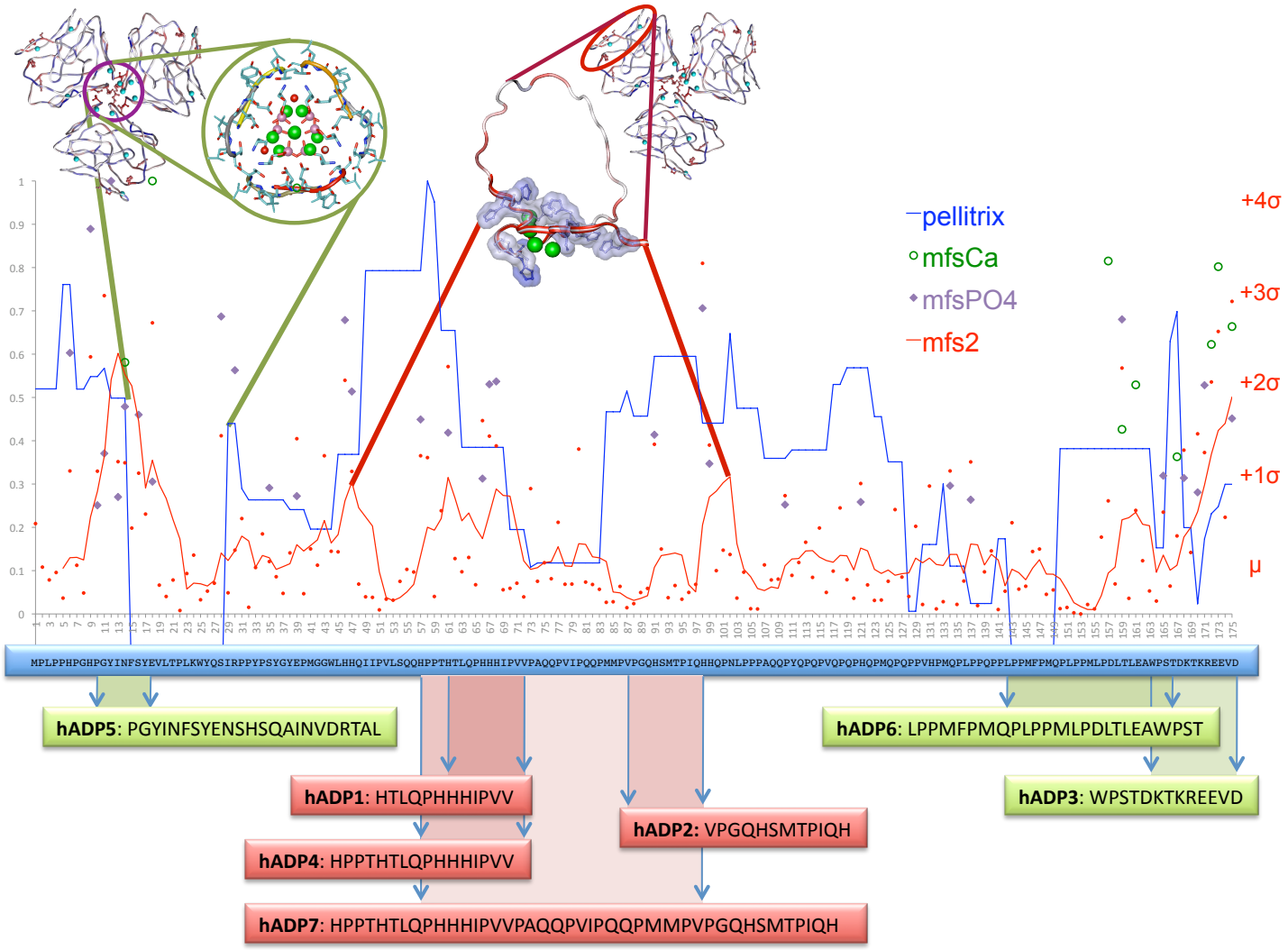

Figure 9. Summary of novel AMELX functionality. Computational prediction of AMELX regions responsible for nucleation and maturation of hydroxyapatite. The YINFSYE region (top left in green) was predicted to nucleate $CaPO_4$ by functional importance, structure- and sequence-based calcium ion binding, phosphate binding, negative enamel binding, quaternary structure and rebuilding around the hydroxyapatite unit cell (circular inset). The HΦ region (top center in red) was predicted by functional importance, phosphate, calcium ion, and enamel binding, and tertiary structure to bind and mature hydroxyapatite. Data plot as described in figure 6. At bottom are the human sequences (hADP) corresponding to the murine AMELX-derived peptides currently undergoing study.

Putative functional regions				
1-32	A domain; NFF	***Yellow***	*Tube*	(Paine and Snead, 1997)
33-46	ATMP; NFF	***Orange***	*Tube*	(Ravindranath et al., 1999)
150-163	B domain; NFF	***Yellow***	*Tube*	(Paine and Snead, 1997)
167-175	COOH-terminal	***Green***	*Tube*	(Aoba et al., 1992)
Putative functional residues				
16	PO_4; NFF	***Pink***	*Sphere*	(Fincham et al., 1994)
Amelogenesis Imperfecta				
21	AI: T→I; NFF	***MFS***	*Surface*	(Lench and Winter, 1995)
22	AI: P→R; NFF	***MFS***	*Surface*	(Kida et al., 2007)
40	AI: P→T; NFF	***MFS***	*Surface*	(Collier et al., 1997)
46	AI: H→L; NFF	***MFS***	*Surface*	(Hart et al., 2002)
Surface exposed residues				
25	Tryptophan	***Purple***	*Stick*	(Oobatake et al., 2006)
44	Tryptophan	***Purple***	*Stick*	"
163	Tryptophan	***Purple***	*Stick*	"
Bonds cleaved by MMP-20 in both rP172 and rM179				
44-45	MMP-20 → TRAP; NFF	***Cyan****	*Tube*	(Ryu et al., 1999)
134-135	MMP-20	***Cyan****	*Tube*	"
146-148	MMP-20 → 146; NFF	***Cyan****	*Tube*	"
Bonds cleaved by non-AMELX-specific endoproteases				
18-19	Glu-C	***Black****	*Tube*	(Moradian-Oldak et al., 2001)
24-25	Trypsin	***Black****	*Tube*	"
25-26	Chymotrypsin	***Black****	*Tube*	"
29-30	Thermolysin	***Black****	*Tube*	"
30-31	Trypsin	***Black****	*Tube*	"
62-63	Thermolysin	***Black****	*Tube*	"
111-112	Chymotrypsin	***Black****	*Tube*	"
146-148	Chymotrypsin†	***Black****	*Tube*	"
153-154	Thermolysin	***Black****	*Tube*	"
167-168	Trypsin	***Black****	*Tube*	"
173-174	Glu-C	***Black****	*Tube*	"
Top 15% Meta-functional signature scores				
Various	MFS scores	***Blue***	*Stick*	(Wang et al., *2008*)
Calcium ion binding site prediction				
NA	SOAK method	***Green***	*Sphere*	(Cheng et al., *in preparation*)

* Representations are partially transparent.

† Preferentially cleaved by MMP-20 and Thermolysin, from many possible sites throughout AMELX.

Table 1. AMELX sequence with annotations from experimental data. Putative functional regions: A & B domains are involved in AMELX-AMELX interactions (Paine and Snead, 1997); ATMP is necessary for cell surface interactions (Ravindranath et al., 2003); the COOH-terminal is thought to interact with the HAp surface (Aoba et al., 1992). Putative functional residues: The phosphorylation of serine-16 is relevant to function (Fincham et al., 1994). Methyl 13C-labeled Ala46 in M59 (corresponding to Ser165 in H175) is 8.0 Ångstroms from the HAp surface (Shaw et al., 2004). Amelogenesis imperfecta: Single residue mutations p.T21I (Lench and Winter, 1995), p.P40T (Collier et al., 1997), p.H46L (Hart et al., 2002), p.P22R (Kida et al., 2007), cause AI. The proteolysis of specific bonds by non-specific and specific endoproteases demonstrates surface presentation for these bonds (Moradian-Oldak et al., 2001; Ryu et al., 1999). Three tryptophans were shown to be solvent accessible (Oobatake et al., 2006).

MFS score (Z-score)	Alteration	Phenotype	Reference
0.84 (2.22)	p.S16 + PO_4	Unknown	(Fincham et al., 1994)
0.57 (0.63)	AI-related p.T21I	Hypomaturation	(Lench and Winter, 1995)
0.68 (1.28)	AI-related p.P22R	Smooth hypoplastic	(Kida et al., 2007)
0.62 (0.92)	AI-related p.P40T	Hypomaturation	(Collier et al., 1997)
0.76 (1.75)	AI-related p.H46L	Hypomaturation	(Hart et al., 2002)
0.28 (-1.07)	p.P169K, p.P169T	No effect	(Paine et al., 2003)

Table 2. MFS scores for known functionally important AMELX residues. MFS scores for five residues specifically known to be critical to AMELX function and one not affecting function are displayed to provide a gauge for the interpretation of MFS score predictions. MFS scores are between 0.00 and 1.00 with the latter indicating maximal probability of contribution to function. Z-scores are given to demonstrate placement of these scores within the distribution of MFS scores for AMELX.

Mean MFS (Z-score)	Max	Region	Residues	Sequence
Previously hypothesized functional regions of AMELX				
0.46 (NA)	1.00	**AMELX All**	1-175	-
0.46 (0.01)	0.84	**A domain**	1-33	MPLPPHPG**HP**GYINF**S** Y**E**VLT**P**L**K**WYQ**S**IRPPY
0.54 (0.42)	0.78	**ATMP**	33-44	Y**P**SYG**YE**PMGGW
0.48 (0.11)	1.00	**B domain**	150-163	PLP**P**MLP**D**LTL**E**AW
0.48 (0.10)	0.57	**carboxy-terminal**	167-175	DKTKREEVD
Novel hypothesized functional regions of AMELX				
0.58 (0.69)	0.84	**Distal A domain**	16-28	**S**Y**E**VLT**P**L**K**WYQ**S**
0.57 (0.63)	0.83	**Proximal HΦ**	54-68	SQQ**H**PPT**H**TLQP**HHH**
0.58 (0.65)	0.94	**Central HΦ**	91-99	**HS**MTPIQ**HH**
0.46 (0.03)	0.62	**Distal HΦ**	121-135	HQPMQPQPPVHPMQP
Novel hypothesized functional regions of LRAP ± 4				
0.39 (NA)	1.00	**LRAP-4 All**	1-59	-
0.69 (1.21)	1.00	**Distal A domain**	12-26	**Y**I**N**L**SYE**VL**TP**L**KWY**
0.41 (NA)	1.00	**LRAP+4 All**	1-73	-
0.73 (1.34)	1.00	**Central A**	12-19	**Y**I**N**L**SYE**K
0.37 (-0.15)	0.56	**Exon 4**	19-32	KSHSQAINTDRTAL
0.65 (1.01)	0.99	**Distal A**	33-40	VL**T**PL**KWY**

Table 3. MFS scores by AMELX region. Mean, standard deviation, and maximum MFS scores are shown for putative human AMELX and murine LRAP functional regions. Residues scored within the top 15% of all MFS scores for each protein are underlined.