Mapping the polymerization landscape of human serpins across genetic variation using Gaussian Processes

Authors: Shuyuan Chen[1,2], Yunsong Deng[2,3,4], Xin Zhang[1#] and Chao Wang[2,3#*]

Affiliations: [1]Department of Chemistry, Westlake University, Hangzhou 310030, China. [2]Institute of Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen, 518132, China. [3]Shenzhen Medical Academy of Research and Translation (SMART), Shenzhen 518107, China. [4]Westlake University, Hangzhou 310030, China.

[#]Correspondence: chaowang@szbl.ac.cn; zhangxin@westlake.edu.cn

[*]Lead contact: chaowang@szbl.ac.cn

Abstract

Proteins function on rugged energy landscapes that enable conformational plasticity but increase vulnerability to misfolding under genetic variation and stress. Serine protease inhibitors (serpins) are metastable proteins whose variants can polymerize and aggregate, causing diseases including emphysema, angioedema, thrombosis and dementia. Existing aggregation predictors underperform on serpin variants. Here, we present a Gaussian process (GP) framework with a sequence position-dependent spatial covariance (SCV) kernel that integrates residue position and structure-informed stability perturbations (ΔΔG) to infer residue-resolved polymerization landscapes across human serpins, together with posterior uncertainty. The model supports a stability-dependent, C-terminal polymerization mechanism across serpin family members and predicts previously uncharacterized human serpin variants with high polymerogenicity. Experimental testing in cells confirmed increased intracellular polymer burden for high-risk variants. By capturing this common, stability-dependent C-terminal polymerization mechanism based on genetic variation in the population, the GP-SCV framework suggests a potential unified strategy for variant prioritization and therapeutic development in serpinopathies.

Introduction

Proteins fold into their native conformations to perform biological functions. To meet functional demands within a dynamic cellular environment, evolutionary pressure often selects sequences capable of sampling multiple conformations[1-3]. These alternative states correspond to local free energy minima on a rugged energy landscape, whose populations are shifted by factors like ligand or substrate binding, allosteric regulation, post-translational modifications and protein–protein interactions[4]. While this roughness is essential for functional plasticity, it also presents an inherent risk for protein misfolding and aggregation, particularly under genetic mutations or environmental stresses such as aging[5,6]. Such misfolding events can trigger cascades of cellular dysfunction that are associated with various systemic and neurodegenerative diseases[7].

Alpha-1 antitrypsin (AAT), encoded by the *SERPINA1* gene, is a metastable protein and the archetypal member of the serine protease inhibitor (serpin) family[8]. AAT is primarily synthesized and secreted by hepatocytes at grams per day into the bloodstream, from which it reaches the lung to neutralize human neutrophil elastase (NE) and protects lung tissue from NE-mediated degradation (**Fig. 1a**, top)[8]. This inhibitory mechanism is initiated when NE cleaves the reactive center loop (RCL) of AAT, triggering the insertion of the cleaved RCL into β-sheet A and translocation of the bound protease, which effectively traps NE in an inactive state (**Fig. 1b**)[9]. The extensive conformational changes required for AAT function also render it prone to misfolding (**Fig. 1c, d**)[1,10-13]. Genetic variants that favor polymerization and aggregation lead to retention of AAT within the endoplasmic reticulum (ER) of hepatocytes and the development of liver disease phenotypes such as neonatal hepatitis, jaundice, fibrosis and cirrhosis (**Fig. 1a**, bottom)[14,15].

Consequently, aggregation in the liver reduces the secretion of functional AAT into the plasma, leading to a loss-of-function in the lung and clinically manifesting as emphysema, bronchitis, and chronic obstructive pulmonary disease (COPD) (**Fig. 1a**, bottom)[8,16,17].

AAT deficiency (AATD) is among the most common inherited disorders with an estimated prevalence of ~1 in 2,000 in non-Finnish European populations. In this population, the severe Z allele (c.1096G>A; p.Glu366Lys) has an estimated allele frequency of ~1.9%, whereas the milder S allele (c.863A>T; p.Glu288Val) occurs at ~4.4%[18]. AATD also occurs on other continents, albeit at lower frequencies[19]. Beyond the common Z and S alleles, approximately 100 rare *SERPINA1* variants have been linked to clinical phenotypes[20-22], among >600 variants observed in population sequencing[18]. AATD therefore provides a paradigmatic model of conformational disease that links folding dynamics, polymerization mechanisms, and disease-causing variation[23].

The tendency toward misfolding and aggregation is not unique to AAT but represents a broader vulnerability of the serpin fold[24]. Mutations in other serpin family members can also lead to protein misfolding and disease phenotypes. For example, familial encephalopathy with neuroserpin inclusion bodies (FENIB) results from the accumulation of neuroserpin (*SERPINI1*) polymers within neurons, which causes progressive symptoms including cognitive decline, tremors, and seizures, ultimately leading to pre-senile dementia and death[25]. Pathogenic mutations in the *SERPING1* gene, which encodes plasma protein C1 inhibitor, lead to hereditary angioedema[26]. Large-scale sequencing of the human genome has uncovered many rare serpin variants of uncertain significance that potentially contribute to diseases through protein misfolding and aggregation[18,20-22,27,28]. Consequently,

it is essential to develop a method to understand and predict the polymer burden of these variants found in the population.

Several computational algorithms, such as TANGO[29] and AggreScan3D[30], have been developed to predict protein aggregation. However, these tools were mostly designed to assess the aggregation propensity of amyloid, which typically adopts a characteristic cross-β sheet structure[31,32]. In contrast, serpin polymers exhibit a distinct "beads-on-a-string" morphology by electron microscopy[33]. Various models have been proposed to explain the polymerization of AAT, including the loop-sheet, β-hairpin, and C-terminal mechanisms[12] (**Fig. 1d**). Converging evidence from structural analyses of patient-derived AAT polymers[10,11,13,33,34], recurrent somatic C-terminal truncations in patient tissues[35], and machine-learning analyses of population variant effects[36-38] supports a C-terminal domain-swap polymerization mechanism for AAT. Whether this sequential intermolecular donation of one molecule's C-terminus into the next predominates across the serpin family remains unresolved, limiting reliable prediction of polymerogenic variants in serpins.

At present, the only disease-specific treatment for AATD in clinical use is intravenous AAT augmentation, which is costly, labor-intensive, and of debated clinical benefit[39,40]. Small-molecule pharmacological chaperones such as GSK716 have been developed to suppress polymerization of AAT mutants by binding a cryptic pocket near the Z-variant site, stabilizing a monomeric folding intermediate and thereby increasing secretion of monomeric AAT[41,42]. However, this mode of stabilization compromises AAT's antiprotease activity[41], underscoring the need for alternative mechanisms that prevent polymerization while preserving protease inhibitory function. The folding, aggregation, and clearance of AAT mutants in the ER of cells are also shaped by a proteostasis network[43-45]. We recently

developed a Gaussian process (GP)-based machine learning framework to model variant effects under targeted proteostasis perturbations[36-38]. Using this approach, we identified GRP94, the ER Hsp90 paralogue, as a regulator of long-range interactions that suppresses C-terminal domain-swap polymerization and improves secretion of functional monomeric AAT[37]. These findings support alternative strategies that target C-terminal domain-swap polymerization while preserving protease inhibitory function[36,37]. A deeper understanding of polymerization mechanisms across the serpin fold could accelerate therapeutic development for serpinopathies.

Here, we further developed the GP-based framework by linking variant sequence position with structural stability changes to predict polymerization propensity across serpin variants. A GP prior defined by a spatial covariance (SCV) kernel is conditioned on a sparse set of experimentally measured variants. This yields a posterior that generates comprehensive, residue-wise polymerization landscapes with quantified uncertainty. These results support a stability-dependent, C-terminal domain-swap polymerization mechanism across multiple serpin family members and identify previously uncharacterized serpin variants that form intracellular aggregates. We pinpoint key sequence and structural regions responsible for polymerization, providing a potential common mechanism for managing serpinopathies.

Results

Existing aggregation prediction algorithms perform poorly on AAT variants

To assess whether existing aggregation prediction algorithms are capable of predicting the polymerization propensity of serpin variants, we utilized 68 AAT missense variants,

whose intracellular polymerization levels have been characterized by the AAT polymer-specific antibody 2C1[37]. The locations of these variants span across the entire AAT sequence (**Fig. 2a**) and 3D structure (**Fig. 2b**). They display a broad range of intracellular polymer burden in hepatocyte-derived Huh7.5 cells in which endogenous AAT was silenced by CRISPR–Cas9[46,47] (**Fig. 2c**), indicating that distinct mutations elicit different structural perturbations that drive polymer formation to varying extents.

We applied three aggregation predictors, TANGO[29], AggreScan3D[30] and ANuPP[48] to estimate the polymerization propensity of AAT variants. None of the predictor scores correlated with experimentally measured polymer burden (**Fig. 2d, e**), indicating that existing aggregation predictors do not capture the determinants of AAT polymerization. Because destabilizing variants can shift the folding landscape toward off-pathway intermediates[49], we next asked whether stability predictors are more informative. Structure-based ΔΔG estimates from FoldX[50], which uses an empirical energy function to quantify variant-induced changes in folding free energy, were weakly associated with AAT polymer burden (**Fig. 2d, e**; Pearson's r = 0.29; Spearman's ρ = 0.23). We also evaluated ThermoMPNN[51], a ProteinMPNN-derived deep-learning model trained on stability datasets to predict variant ΔΔG. Similar to FoldX, ThermoMPNN showed weak correlations with experimentally measured polymer burden (**Fig. 2d, e**; Pearson's r = 0.31; Spearman's ρ = 0.32). Together, these results indicate that variant-induced stability changes explain only a limited component of AAT polymerization propensity and are insufficient to predict polymer burden accurately.

Mapping the stability-dependent AAT polymerization landscape using GP

We previously developed a GP-based machine learning framework to show that AAT polymerization is governed by discrete sequence regions[36-38]. To understand where and how structural stability modulates polymerization, we adapted the GP framework by using an SCV kernel that defines covariance in polymer burden between variants based on their proximity in a joint space of normalized sequence position and structure-informed stability perturbations. GPs provide a powerful framework for Bayesian nonparametric modeling in which the kernel specifies a prior over functions, and conditioning on sparse measurements yields a posterior predictive distribution that provides both point estimates and uncertainty for unmeasured variants (**Fig. 3a**).

To construct a stability-dependent polymerization landscape along the AAT sequence, we normalized each variant's residue position to the length of the mature AAT polypeptide after removing the signal peptide so that positions range from 0 to 1, with 1 denoting the C-terminus (**Fig. 3b, *x* axis**). Predicted structural stability perturbations (ΔΔG) and experimentally measured polymer burden were scaled to the interval from 0 to 1 using the minimum and maximum values and plotted on the *y* and *z* axes, respectively (**Fig. 3b**). We then computed distances between variants in the joint space of normalized position and stability (**Fig. 3b,** black lines) and used these distances to define covariance in polymer burden, thereby specifying the SCV kernel. The implied covariance structure can be visualized with a variogram, which plots semivariance in polymer burden as a function of distance in the joint feature space (**Supplementary Fig. 1a**). Hyperparameters were chosen by balancing variogram fit and predictive performance to obtain a kernel that captures the global covariance structure while maintaining robust predictive accuracy (**Supplementary**

Fig. 1a-d; see Methods). Using the optimized kernel, the GP generates a residue-resolved polymerization landscape with quantified uncertainty along the AAT sequence (**Fig. 3c**).

We built GP models using stability perturbations predicted by either FoldX or ThermoMPNN as inputs, referred to as GP-FoldX and GP-ThermoMPNN, respectively (**Fig. 3c-e**; **Supplementary Fig. 1e**). In leave-one-out cross validation, both GP models yielded higher Pearson and Spearman correlations with measured polymer burden than using ΔΔG as a direct baseline predictor (**Fig. 3d, e**). GP-based models also outperformed other machine-learning approaches, including random forests, XGBoost and neural networks, when evaluated using the same input features and leave-one-out protocol (**Supplementary Fig. 1f**), consistent with the suitability of GPs for sparse-data settings[37,52-55]. Together, these results highlight a position-dependent coupling between stability perturbations and polymer burden and motivate the use of GP to map polymerization landscapes across the AAT polypeptide chain.

GP uncertainty informs residue-level stability effects on AAT polymerization

The stability-dependent polymerization landscape generated with the SCV kernel maps how polymer burden varies across the joint space of sequence position and predicted ΔΔG (**Fig. 3c**, **Supplementary Fig. 1e**). Because the GP provides an uncertainty estimate for each prediction, we classified points across the landscape into high- and low-confidence groups using a Gaussian mixture model (**Supplementary Fig. 1g, h**). High-confidence regions were delineated by contour lines, covering ~96.7% of residues in the mature AAT sequence, whereas only a short N-terminal fragment is absent from the available structure (PDB: 3NE4), precluding ΔΔG estimation and thus high-confidence prediction (**Fig. 3c; Supplementary Fig. 1e**). Notably, the landscape highlights prominent high-confidence

regions near the AAT C-terminus, where destabilizing variants are most strongly associated with increased polymer burden (**Fig. 3c; Supplementary Fig. 1e**).

To quantify residue-level contributions to stability-dependent AAT polymerization, we summarized predictions for each residue across the range of ΔΔG inputs by inverse variance weighting[36-38], using GP predictive uncertainty as weights (**Fig. 3c**; see Methods). This yields a residue-resolved, uncertainty-weighted mean polymer burden score that marginalizes over the ΔΔG dimension (**Fig. 3f**). The scores indicate that stability perturbations in the C-terminal half of AAT are more strongly associated with polymer formation, with the strongest signal at the C-terminal s5B β-strand. Most positions in the N-terminal half show lower scores, although the N-terminal *s6B* β-strand displays a localized elevation in scores (**Fig. 3f**). We mapped these residue scores onto the AAT monomer[56] (**Fig. 3g**) and polymer[57] structures (**Fig. 3h**). The mapping highlights regions involved in 'loop-sheet' insertion, including the RCL and s5A, where stability perturbations are associated with polymer formation (**Fig. 3g, h**). It also indicates β-sheet B, particularly *s3B*, *s4B*, and *s5B* β-strands, as a major stability-sensitive determinant of polymerization. These patterns are consistent with a C-terminal domain swap route and suggest that population variants can promote AAT polymer formation by perturbing β-sheet B stability.

Profiling polymerogenicity of AAT variants in human populations using GP

Large-scale sequencing has revealed extensive rare coding variation in human populations[18]. In *SERPINA1,* beyond the well-studied Z and S alleles, approximately 600 missense variants have been reported[18,21,22], the vast majority of which remain functionally uncharacterized. Building on our stability-dependent GP framework, we predicted polymerogenicity for AAT missense variants in gnomAD[18] (**Fig. 4a**). The predicted

polymerogenicity distribution is unimodal and right-skewed, with most variants assigned low predicted polymerogenicity and a minority extending into a long high-polymerogenicity tail (**Fig. 4a**). Using the Z-allele leave-one-out posterior mean as a reference threshold (**Fig. 4a**, red dashed line), a subset of population variants exceeded this cutoff and were prioritized as candidates with Z-like or higher predicted propensity to form intracellular polymers.

To understand how population variation samples the stability-dependent polymerization landscape along AAT sequence, we averaged GP predictions across variants observed at each residue position (**Fig. 4b**). Variants were distributed across most of the polypeptide, including regions with elevated predicted polymerogenicity (**Fig. 4b**). We further visualized variant composition by plotting per-position variant fractions for the first 50 N-terminal and last 150 C-terminal residues, coloring each substitution by its predicted polymerogenicity (**Fig. 4c**). Consistent with the residue-level profile (**Fig. 3f**), human AAT variants in the N-terminal half generally show lower predicted polymerization propensity (**Fig. 4b**), with notable exceptions concentrated in the *hA–s6B–hB* region, exemplified by I74N[58-61] and V79E[62] (**Fig. 4c**). Notably, V79E was not included in the training set yet has recently been reported in individuals with AATD[62], indicating that the model can prioritize previously unmeasured variants with clinical relevance. Overall, variants in the C-terminal half exhibited higher predicted polymerogenicity and clustered into three hotspots (**Fig. 4b, c**). The first lies near the S allele (E288V) region and includes variants such as L278P and L287P[63] with strong predicted polymerization signals (**Fig. 4b, c**). A second hotspot spans *s5A* (residues 358–362), immediately preceding the RCL, a segment required for loop–sheet insertion (**Fig. 4b, c**). The third hotspot maps to the C-

terminal *s4B–s5B* strands that mediate C-terminal donation during domain-swap polymerization (**Fig. 4b, c**), suggesting that human population variation samples residues that are mechanistically central to polymer formation.

To assess the clinical relevance of the population-scale predictions, we compared predicted polymerogenicity for AAT missense variants annotated in ClinVar as benign/likely benign versus pathogenic/likely pathogenic. Variants labeled pathogenic/likely pathogenic showed significantly higher predicted polymerogenicity than benign/likely benign variants (**Fig. 4d**), consistent with their clinical classification. Collectively, our GP-based profiling delineates polymerogenicity across population variants and provides a principled route to prioritize candidates for experimental follow-up.

Experimental validation identifies previously uncharacterized polymerogenic AAT variants

We next prioritized variants for experimental validation based on GP models using two independent ΔΔG estimators, FoldX and ThermoMPNN. Across population variants not included among experimentally measured training variants, GP-FoldX and GP-ThermoMPNN predictions were moderately correlated (**Fig. 4e**; Pearson's r = 0.45). We then defined model-specific Z reference thresholds using the leave-one-out posterior mean for the Z allele (**Fig. 4e**; red dashed lines) and selected variants exceeding this reference in both models (**Fig. 4e**; red points). Seventeen AAT variants met these criteria and were carried forward for cellular assays of intracellular polymer burden.

We generated expression plasmids encoding HA-tagged AAT variants together with an IRES–GFP reporter to monitor transfection and expression. Plasmids were transiently expressed in Huh7 cells, and both cell lysates and conditioned media were collected to quantify intracellular and secreted AAT. Lysates were fractionated into soluble and insoluble fractions, and insoluble AAT was recovered for immunoblot analysis as a readout of intracellular polymer burden. As expected, the Z and S variants showed reduced secretion and increased intracellular insoluble AAT compared with AAT-WT (**Fig. 4f-h**), consistent with their established polymerization phenotypes.

Among the seventeen prioritized variants, thirteen (76%) produced a significant insoluble AAT signal (**Fig. 4f-h**), indicating intracellular accumulation of insoluble AAT consistent with polymer formation. Several of these polymerogenic variants, including V79E[62], A360P[64], D365H[65] and M409T[22], have been reported in clinical settings, supporting the utility of the model for nominating clinically relevant polymer-forming candidates. The remaining polymerogenic variants have not, to our knowledge, been reported clinically, but are present in population sequencing resources[18] (**Fig. 4f-h**). These variants therefore represent candidate disease-associated alleles that warrant further investigation. Together, these validation experiments support the GP-based prioritization strategy and motivate extending the framework to map polymerogenicity across the broader serpin family.

Stability-dependent GP polymerization landscape across human serpins

Humans encode 36 serpin proteins, with additional serpin pseudogenes (**Fig. 5a**). Serpins share a conserved fold built around three β-sheets and surrounding α-helices, together with a solvent-exposed RCL that undergoes large conformational rearrangements

during function (**Fig. 2b**; **Supplementary Fig. 2**). Across the family, the α/β core is highly conserved, whereas terminal extensions and surface loops are more variable and are frequently assigned lower confidence by AlphaFold2[66], consistent with protein-specific flexible or disordered regions (**Supplementary Fig. 2**). Population sequencing shows that each serpin member harbors hundreds of coding variants in gnomAD (**Fig. 5b**). Given the shared structural core of the serpin fold, we reasoned that the stability-dependent GP framework developed for AAT could be generalized to prioritize polymerogenic variants across the human serpin family.

As AAT is the archetypal serpin, we aligned each serpin family member to AAT to map variant residues onto an AAT-aligned position, which served as the positional input to the model (**Fig. 5c**). For each variant, we estimated stability perturbations (ΔΔG) using either FoldX or ThermoMPNN on AlphaFold2-predicted serpin structures (**Fig. 5c**; **Supplementary Fig. 2**). These features were then used to predict variant polymerogenicity with the GP model trained on experimentally measured AAT polymer burden (**Fig. 5c**).

To assess correspondence with clinical interpretation for serpin family members, we compared predictions to ClinVar germline classifications. We focused on serpins with at least three variants annotated in each of the benign/likely benign and pathogenic/likely pathogenic categories, yielding four genes, SERPING1, SERPINI1, SERPINC1, and SERPINA7. For SERPING1, SERPINI1 and SERPINC1, pathogenic/likely pathogenic variants showed significantly higher predicted polymerogenicity than benign/likely benign variants (**Fig. 5d**), consistent with polymerization-related mechanisms contributing to disease in these serpins. For SERPINA7, the two distributions overlapped (**Fig. 5d**), which

311 may in part reflect the limited number of ClinVar-classified variants available for
312 comparison (**Fig. 5d**).

313 We generated predictions of polymerogenicity for missense variants observed in
314 gnomAD for each serpin gene (**Fig. 5e**). Each serpin contained variants exceeding the Z-
315 allele leave-one-out posterior mean, flagging candidates that may be at elevated risk of
316 polymer formation (**Fig. 5e**). The overall prediction levels differed across family members,
317 with some serpins showing higher mean polymerogenicity than others. To examine where
318 population variation samples polymerization-prone regions of the fold, we averaged
319 predictions across variants observed at each residue position and plotted residue-resolved
320 polymerogenicity (**Fig. 5f**). This analysis revealed a broadly shared positional pattern
321 across serpins, with recurrent hotspots that are most prominent toward the C-terminal
322 portion of the fold (**Fig. 5f**). The strongest signal concentrates in the C-terminal β-sheet B
323 region, consistent with a C-terminal domain-swap polymerization mechanism shared
324 across serpin family members (**Fig. 5f**).

325 **Experimental validation of polymerogenic variants in SERPINI1**

326 To experimentally test polymerogenicity predictions across serpin family members,
327 we focused on neuroserpin (*SERPINI1*), whose pathogenic variants cause FENIB, a
328 neurodegenerative disorder characterized by progressive dementia. Across neuroserpin
329 missense variants, polymerogenicity scores from GP-FoldX and GP–ThermoMPNN were
330 positively correlated (**Fig. 6a**; Pearson's r = 0.56), indicating agreement in variant risk
331 ranking between the two predictors. We selected neuroserpin missense variants with
332 predicted polymerogenicity exceeding the Z-allele leave-one-out posterior mean for
333 experimental validation.

HA-tagged neuroserpin variants were expressed together with an IRES–GFP reporter in HeLa cells, and neuroserpin was measured in the secreted fraction (medium), soluble lysate (supernatant), and insoluble pellet, with GFP and GAPDH providing controls for expression and loading (**Fig. 6b–d**). WT neuroserpin remained predominantly soluble with minimal pellet signal, whereas the established FENIB variants S52R and H338R robustly accumulated in the insoluble fraction, consistent with their known propensity to form intracellular polymers (**Fig. 6b-e**).

Among 26 tested variants, 20 (76.9%) increased insoluble neuroserpin relative to WT (**Fig. 6b-e**), indicating that model-prioritized variants are enriched for elevated insoluble aggregation. Structural mapping showed that variants with high aggregation burden are not uniformly distributed, but instead cluster around β-sheet B and adjacent elements (**Fig. 6f**). Notably, all tested variants with insoluble signal exceeding 80% of the S52R level (**Fig. 6b-e**) localized to the N-terminal *hA-s6B* region (R38/I46) and C-terminal *s4B-s5B* region (F377/I380/F390/G392/R393) (**Fig. 6f**), suggesting that long-range N-to-C coupling that controls β-sheet B stability and accessibility is critical for neuroserpin polymerization and aggregation.

Discussion

Serpins are metastable proteins shaped by a fundamental evolutionary trade-off[1,12,67]. The conformational plasticity required for their inhibitory cycle also increases susceptibility to off-pathway misfolding that promotes polymer formation and aggregation, a common pathogenic mechanism underlying serpinopathies. Naturally occurring genetic variation in human populations samples this conformational landscape and yields a broad spectrum of polymerogenicity. Despite decades of mechanistic work on archetypal

357 members such as AAT, predicting the polymerogenicity for the vast genetic variation in
358 serpins has remained difficult. This is in part because canonical aggregation predictors
359 including TANGO[29], AggreScan3D[30] and ANuPP[48] tested here, were developed primarily
360 for amyloid-like, cross-β aggregation pathways and do not encode the topology and kinetic
361 traps that govern serpin polymerization. Here we address this gap by coupling sparse
362 experimental measurements to a GP framework whose SCV kernel integrates sequence
363 positional information with structure-informed stability perturbations (**Fig. 6g**). This
364 approach yields residue-resolved polymerization landscapes with calibrated uncertainty,
365 enables population-scale variant prioritization, and reveals a mechanistically interpretable
366 link between local destabilization and a recurring polymerization route across the serpin
367 fold (**Fig. 6g**).

368 Although protein stability is known to contribute to aggregation, we found that both
369 empirical force field and deep learning–based stability predictors, including FoldX[50] and
370 ThermoMPNN[51], correlate weakly with measured cellular polymer burden of AAT variants.
371 This is consistent with our previous observation that some AAT variants primarily
372 accelerate monomer degradation without forming detectable polymers in cells, and that
373 polymerization risk is concentrated in specific structural regions rather than distributed
374 uniformly across the fold[36-38]. The GP-SCV model formalizes this intuition by learning a
375 position-dependent coupling between stability perturbations and polymer burden, thereby
376 mapping a residue-resolved polymerization landscape. It shows that identical ΔΔG values
377 can have markedly different polymerization outcomes depending on where in the fold they
378 occur (**Fig. 3c**). The model also provides posterior uncertainty that separates regions tightly
379 constrained by training measurements from regions of lower confidence, thereby

pinpointing residues that most strongly shape AAT polymerization and illuminating the underlying mechanism.

Trained on clinically observed AAT variants, the GP-SCV model indicates that residues where destabilizing perturbations are most strongly associated with polymer burden cluster not only in the canonical *s5A*–RCL region implicated in loop–sheet insertion, but also in stability-sensitive elements of β-sheet B, consistent with a C-terminal domain-swap mechanism. One plausible model is that destabilization of β-sheet B increases the likelihood of *s4B–s5B* disengagement from the native fold, which could facilitate RCL insertion as a new s4A and enable intermolecular insertion of *s4B–s5B* into the β-sheet of a neighboring molecule. Structural details of C-terminal domain swapping are now emerging from *in vivo* and *ex vivo* structures of AAT polymers[10,11,13,33,34], and our residue-resolved landscape aligns with this framework by nominating the C-terminal strands of β-sheet B as a focal point where destabilizing variation most strongly associates with polymer burden. Notably, variants drawn from population genetic datasets that are predicted to be highly polymerogenic also cluster into discrete hotspots that map onto mechanistically interpretable regions, consistent with the idea that human genetic variation samples polymerization-relevant states of the serpin fold[35]. Together, these findings connect population genetics to a residue-level mechanism and provide a principled basis for prioritizing uncharacterized variants for experimental follow-up validation.

Importantly, we find that the same stability-dependent logic extends beyond AAT to other serpin family members, whose variants are known to cause a broad spectrum of disorders collectively termed serpinopathies[24], including emphysema, hereditary angioedema, thrombosis and dementia. By aligning serpins to a common positional

coordinate and estimating stability perturbations on predicted structures, we generated polymerogenicity profiles across the human serpin family that show recurrent enrichment toward the C-terminal portion of the conserved fold, with the strongest shared signal in the β-sheet B region. Notably, across multiple serpin family members, an additional contributing segment is observed near the N-terminal *s6B* region (**Fig. 5f**). Because *s6B* forms part of β-sheet B and interacts with the C-terminal *s5B* strand, this pattern highlights the importance of N-to-C coupling in maintaining serpin fold integrity, and suggests that perturbation of this interaction can strongly influence polymerization propensity.

We experimentally evaluated this framework using AAT and neuroserpin variants. Notably, although no neuroserpin variants were included in model training, 20 of 26 GP-SCV prioritized variants (76.9%) formed significantly increased insoluble aggregates in cells. The strongest polymerogenic variants localized predominantly to the β-sheet B region, supporting the broader conclusion that a stability-dependent C-terminal domain-swap mechanism is not unique to AAT, but may represent a recurring route by which the serpin fold accesses polymeric states. More broadly, the long tail of highly polymerogenic rare missense variants across serpin family members suggests that many individuals may harbor alleles with previously unrecognized polymer risk, warranting further investigation.

Several limitations point to important directions for future work. First, our stability-dependent landscapes rely on predicted stability perturbations rather than direct experimental measurements. Although we used multiple predictors to estimate these effects, their accuracy remains limited. FoldX, for example, typically shows only modest correlation with experimentally measured stability[49], and although ThermoMPNN performs better in certain cases[51], its accuracy is likely to decline for proteins that are

poorly represented in the training data. Second, our family-wide predictions depend on structure-informed ΔΔG estimates derived from predicted structures[66,68] and on a shared positional coordinate obtained through sequence alignment. Consequently, uncertainties in poorly resolved regions, protein-specific insertions, and post-translational modifications such as glycosylation may influence polymer risk in ways not captured by the current model. Third, different serpin family members are expressed in distinct tissues and therefore operate within different proteostasis environments[4-7,43]. Differences in folding, degradation and trafficking capacity across cell types are likely to modulate polymer formation beyond what can be explained by intrinsic sequence and structure features alone. A fuller understanding of serpin polymerization will therefore require integrating intrinsic variant effects with protein-specific structural context and cell-type-specific proteostasis constraints[4-7,36-38,43].

Despite these caveats, our results suggest a potentially unifying framework for serpinopathies. Polymerization risk of serpins concentrates at structural elements that govern β-sheet B integrity and C-terminal accessibility, and naturally occurring variation can be used to map these elements with quantified uncertainty. This framework also points to therapeutic strategies that go beyond globally stabilizing the native state. For example, selectively reinforcing the stabilizing network that restrains β-sheet B, or modulating the accessibility of the C-terminal donor segment, could suppress polymerization while preserving the conformational transitions required for inhibitory function, as suggested by GRP94- and ATF6-directed modulation[36,37]. This is particularly important because some polymer-blocking ligands compromise function. More broadly, the integration of population genetics, structure-informed energetics and probabilistic learning provides a

449 generalizable framework for converting sparse variant measurements into mechanistic
450 landscapes (**Fig. 6g**), with the potential to improve both the interpretation of rare alleles
451 and the development of mechanism-guided interventions across protein misfolding
452 diseases.

453 **Methods**

454 **Input training data and feature normalization**

455 The training set comprised 68 AAT missense variants with experimentally measured
456 intracellular polymer burden from prior work[37]. Of the 71 measured missense variants,
457 three located in the N-terminal signal peptide were excluded because these residues are not
458 part of the mature AAT sequence represented in the 3D structure. For the modeling, we
459 removed the 24-residue signal peptide and indexed residues on the mature AAT sequence.
460 The residue position was encoded as $x = r/394$, where r is the residue index in the mature
461 protein (residue 1 to 394), such that $x \in [0,1]$ and $x = 1$ corresponds to the C-terminus.

462 For the GP-FoldX model, variant stability changes (ΔΔG) were computed with FoldX
463 5.0 (Linux version) using the AAT structure (PDB 3NE4)[56] repaired with RepairPDB
464 followed by BuildModel. For the ThermoMPNN-based model, ΔΔG values were computed
465 with ThermoMPNN[51] v1.0.0 using default settings. We defined $\Delta\Delta G = \Delta G_{Var} - \Delta G_{WT}$, such
466 that more positive ΔΔG values indicate greater destabilization by the variant. For each
467 predictor, raw ΔΔG values were min-max normalized as y = (ΔΔG - y_min)/(y_max -
468 y_min). Experimentally measured polymer burden was normalized in the same form, z =
469 (polymer burden - z_min)/(z_max - z_min).

470 **GP modeling with a stability-dependent spatial covariance kernel**

We adapted the GP modeling framework described previously[36-38,52,54,69]. Briefly, normalized polymer burden (z) was modeled as a Gaussian process over the two-dimensional input feature vector $u = (x, y)$, where x is normalized residue position and y is normalized ΔΔG. Prior covariance between variants was defined as a function of their Euclidean distance in this joint feature space, such that variants closer in both position and stability were assigned higher covariance than those farther apart. We implemented this distance–covariance relationship by fitting an empirical variogram, which parameterizes the semivariance $\gamma(h)$ as a function of separation distance h. The covariance at separation distance h, $C(h)$ is given by $C(h) = C(0) - (\gamma(h) - \gamma(0))$, where $C(0)$ is the total variance of the process (the variogram sill) and $\gamma(0)$ captures the nugget effect. Variogram fitting was performed in R using custom workflows adapted from autofitVariogram with the packages sp, gstat, and automap. Empirical variograms were constructed from pairwise semivariance of z, using distance bins spanning 2% to 100% of 40% of the diagonal length of the joint feature space in 2% increments, and adjacent bins were merged until each bin contained at least five point pairs. A spherical variogram model was used for the GP modeling.

Hyperparameter selection and model evaluation

To mitigate overfitting during hyperparameter selection, we optimized variogram hyperparameters by jointly considering leave-one-out cross-validation (LOOCV) predictive performance and the variogram fit, quantified as the sum of squared errors (sser) between the empirical and fitted variogram. The LOOCV was computed with gstat::krige.cv using nmin = 4 and neighborhood sizes nmax from 14 to 30. Predictive performance was summarized with Pearson correlations between held-out predictions and

494 measured polymer burden. Models were grouped into sser bins of width 0.1 standard
495 deviation, and the model with the highest LOOCV Pearson correlation in each bin was
496 selected as the bin representative. We ranked candidate models by sser and examined the
497 LOOCV Pearson correlation as a function of sser. LOOCV Pearson correlation increased
498 rapidly at low sser and then approached a plateau, whereas further increases in sser
499 provided little additional gain (Supplementary Fig. 1). We therefore selected the model at
500 the onset of this plateau (Supplementary Fig. 1; “elbow” point indicated by arrows),
501 prioritizing the smallest sser among models whose LOOCV Pearson correlation was near-
502 saturated, thereby maintaining close agreement between the fitted and empirical
503 variograms and reducing the risk of selecting overly flexible models. Conditioning the GP
504 covariance prior on the 68 measured variants yielded the posterior predictive mean and
505 variance at each queried point on the landscape. In the plotting workflow, the displayed
506 landscape was interpolated on a dense grid spanning x from 0 to 1 in steps of 1/394 and y
507 from 0 to 1 in steps of 0.001, with kriging performed using nmin = 4 and nmax = 16.
508 Alternative regressors were benchmarked in Python using the same *x* and *y* inputs and
509 the same 68-sample training set. Random forests (36 hyperparameter configurations),
510 XGBoost (24 configurations), and multilayer perceptron neural networks (4 configurations)
511 were evaluated under both unscaled and standardized-input preprocessing, with random
512 seed fixed at 42. Performance was summarized by LOOCV Pearson correlation, LOOCV
513 Spearman correlation, root mean squared error, mean absolute error, and R^2. The model
514 with the best performance in each machine learning method was used for comparison. The
515 benchmark environment recorded numpy 2.4.2, pandas 3.0.0, scikit-learn 1.8.0, and
516 xgboost 3.1.3.

Predictive uncertainty classification and residue-level polymerogenicity summary

GP posterior variance was used as the uncertainty measure for each queried point, which was modeled with a two-component Gaussian mixture model using mclust::densityMclust with G = 2. The principal boundary separating low- and high-variance regions (i.e., high- and low-confidence regions) was defined by the intersection of the two fitted mixture-component densities rather than by an arbitrary percentile threshold. For visualization, additional contour lines were drawn from selected offsets around the low-variance component, matching the plotting workflow used for the phenotype landscapes.

Residue-level polymer burden scores were calculated from high-confidence landscape points whose posterior variance fell below the Gaussian mixture model intersection threshold. For each residue, posterior mean predictions were combined by inverse-variance weighting, with $\omega_i = 1/\sigma_i^2$

$$\bar{z} = \frac{\sum_i \omega_i z_i}{\sum_i \omega_i}$$

where z_i is the GP posterior mean and σ_i is the corresponding posterior standard deviation for the i-th queried point. This weighting increases the contribution of more certain predictions and follows the uncertainty-aware residue summarization used in our previous GP-based variant-landscape analyses. Residue scores were then used for sequence plots and structural coloring; for mapping onto the AAT crystal structure, only residues represented in the available structural model were colored.

Generalization across human serpins

For family-wide prediction, the AAT-trained GP model, selected kernel parameters, and *x*-*y* normalization scheme were kept fixed, and we varied only the input features for the target serpin. Each target serpin sequence was aligned to the human AAT reference sequence (SERPINA1; sp|P01009|A1AT_HUMAN) with BLASTP 2.15.0+ using default parameters, and the aligned residue coordinate on AAT was used as *x*. Variant-specific ΔΔG values were then computed on the corresponding target SERPIN protein with FoldX 5.0 or ThermoMPNN v1.0.0. For proteins outside SERPINA1, structure-based ΔΔG calculations used AlphaFold2 models downloaded from the AlphaFold Protein Structure Database. No additional serpin variants were added to the training set or used to retrain the GP.

Selection of variants for experimental validation

Variants were prioritized for cell-based follow-up only if their posterior mean polymerogenicity exceeded the leave-one-out posterior mean of the pathogenic Z allele in both GP-FoldX and GP-ThermoMPNN. This dual-model rule used the Z allele as a fixed reference and required concordant ranking above that reference in both models. The corresponding thresholds were 0.26 for the FoldX-based model and 0.29 for the ThermoMPNN-based model. Applying this rule yielded 17 SERPINA1 variants from the unmeasured population-variant set and 26 SERPINI1 variants for experimental validation.

Experimental validation of SERPIN variants

Reagents and antibodies

The anti-HA monoclonal antibody used for Western blotting was purchased from Cell Signaling Technology (Cat. No. 3724). Antibodies against GAPDH (Cat. No. 50430-2-AP) and GFP (Cat. No. 60004-1-Ig) were obtained from Proteintech.

Expression vectors and cell transfection

AAT variants cloned into pcDNA3.1(+) expression vector were kindly provided by Dr. Xi Wang from Institute of Biophysics, Chinese Academy of Sciences. SERPINI1 (Neuroserpin) variants cloned in pcDNA3.1(+) vector were generated by Tsingke Biotechnology (China). For both constructs, an IRES–GFP cassette was included to monitor transfection efficiency and expression. Plasmid transfection was performed using Lipofectamine 2000 (Thermo Fisher) following the manufacturer's protocol.

Immunoblot analysis

Huh7 cells (purchased from Procell, Cat. No. CL-0120) and HeLa cells (kindly provided by Prof. Yongqiang Deng at Southern Medical University, who purchased the cells from Procell, Cat. No. CL-0101) were transfected with AAT and neuroserpin plasmids, respectively, for 48 h. Cells were washed twice with pre-warmed phosphate-buffered saline (PBS) and subsequently incubated in serum-free medium for 4 h, after which conditioned media were collected. Cells were harvested and lysed in lysis buffer containing 50 mM Tris-HCl (pH 7.4), 150 mM NaCl, and 1% (v/v) Triton X-100 supplemented with a protease inhibitor cocktail (NCM Biotech, P002). The lysates were centrifuged at 14,000 rpm for 10 min at 4°C to separate the supernatant and pellet fractions. The pellet fraction was washed twice with buffer containing 50 mM Tris-HCl (pH 7.4) and 150 mM NaCl, and then resuspended in buffer containing 50 mM Tris-HCl (pH 6.8), 5% SDS, and 10% glycerol. All samples were mixed with SDS loading buffer and heated at 95°C for 5 min. Equal volumes of protein samples were resolved by 10% SDS–PAGE and transferred onto polyvinylidene difluoride (PVDF) membranes. Membranes were blocked with 5% non-fat milk in TBS-Tween 20 (TBST) for 1 h at room temperature, followed by incubation with primary antibodies and horseradish peroxidase (HRP)-conjugated secondary antibodies.

Protein bands were visualized using an enhanced chemiluminescence (ECL) detection system (Beyotime, P0018FM). Band intensities were quantified by densitometry, and insoluble signal was normalized to GFP to derive an intracellular insoluble polymer burden metric.

Statistical analysis

All statistical analyses were performed using R or OriginPro software. Differences in predicted polymerogenicity between ClinVar-annotated benign/likely benign and pathogenic/likely pathogenic variants in serpin genes (**Fig. 4d**, **Fig. 5d**) were assessed using an unpaired two-tailed Welch's t-test. For experimentally measured intracellular polymer burden (**Fig. 4h**, **Fig. 6e**), insoluble serpin polymer signals were quantified from western blots (three independent biological replicates per variant). Within each replicate blot, signals were first normalized to the positive polymer control (Z variant in AAT and S52R in neuroserpin) run on the same blot, with the positive control polymer level set to 1. To test whether a variant exhibited increased polymerization relative to WT, we performed a within-replicate comparison to the WT lane on the same blot using a one-sided paired Student's t-test (paired by replicate; alternative hypothesis: variant > WT). P values across all tested variants were adjusted for multiple comparisons using the Benjamini–Hochberg false discovery rate (BH-FDR) procedure, and FDR-adjusted q values are reported.

Data and code availability

Training data, trained models and prediction scripts will be available on GitHub upon publication. For review purposes, these materials are currently accessible at https://aat-gp-models.pages.dev/96ea19892086fa2caf77a2764522eee70a98a3fd/codes/.

Acknowledgements

We thank Fang Xiao and Jiameng Hu for technical support with GP modeling and figure preparation. Grants were provided by National Key R&D Program of China, 2025YFA0923000 (to C.W.); National Natural Science Foundation of China, 82471903 (to C.W.); Guangdong Provincial Key R&D Program, 2024B1111150001 (to C.W.); Guangdong Pearl River Funding, 2023QN10Y299 (to C.W.); Shenzhen Bay Laboratory (21240111 to C.W.); National Natural Science Foundation of China, 22494700 and 22477102 (to X.Z.); "Pioneer" and "Leading Goose" R&D Program of Zhejiang, 2025SDXHDX000* (to X.Z.); and Westlake Laboratory of Life Sciences and Biomedicine, 2024SSYS0035 (to X.Z.).

Author contributions

S.C. and C.W. performed the computational analysis; Y.D. performed the experimental validation. C.W. conceived and designed the study. C.W. and X.Z. supervised the work and acquired funding. C.W. and S.C. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

References:

1 Gershenson, A., Gierasch, L. M., Pastore, A. & Radford, S. E. Energy landscapes of functional proteins are inherently risky. *Nat Chem Biol* **10**, 884-891 (2014).

2 Campbell, E. C. *et al.* Laboratory evolution of protein conformational dynamics. *Curr Opin Struct Biol* **50**, 49-57 (2018).

3 Hilser, V. J. *et al.* Statistical Thermodynamics of the Protein Ensemble: Mediating Function and Evolution. *Annu Rev Biophys* **54**, 227-247 (2025).

631 4 Powers, E. T. & Gierasch, L. M. The Proteome Folding Problem and Cellular
632 Proteostasis. *Journal of Molecular Biology* **433**, 167197 (2021).
633 5 Hipp, M. S., Kasturi, P. & Hartl, F. U. The proteostasis network and its decline in
634 ageing. *Nat Rev Mol Cell Biol* **20**, 421-435 (2019).
635 6 Hetz, C. & Dillin, A. Central role of the ER proteostasis network in healthy aging.
636 *Trends in Cell Biology* **35**, 548-561 (2025).
637 7 Hipp, M. S. & Hartl, F. U. Interplay of Proteostasis Capacity and Protein
638 Aggregation: Implications for Cellular Function and Disease. *J Mol Biol* **436**, 168615
639 (2024).
640 8 Strnad, P., McElvaney, N. G. & Lomas, D. A. Alpha(1)-Antitrypsin Deficiency. *N*
641 *Engl J Med* **382**, 1443-1455 (2020).
642 9 Huntington, J. A., Read, R. J. & Carrell, R. W. Structure of a serpin-protease
643 complex shows inhibition by deformation. *Nature* **407**, 923-926 (2000).
644 10 Aldobiyan, I. *et al.* The mechanism of pathogenic α(1)-antitrypsin aggregation in
645 the human liver. *Proc Natl Acad Sci U S A* **122**, e2507535122 (2025).
646 11 Lowen, S. M. *et al.* High-resolution characterization of ex vivo AAT polymers by
647 solution-state NMR spectroscopy. *Sci Adv* **11**, eadu7064 (2025).
648 12 Duvoix, A., Roussel, B. D. & Lomas, D. A. Molecular pathogenesis of alpha-1-
649 antitrypsin deficiency. *Rev Mal Respir* **31**, 992-1002 (2014).
650 13 Jagger, A. M., Waudby, C. A., Irving, J. A., Christodoulou, J. & Lomas, D. A. High-
651 resolution ex vivo NMR spectroscopy of human Z α(1)-antitrypsin. *Nat Commun* **11**, 6371
652 (2020).

14 Fromme, M., Schneider, C. V., Trautwein, C., Brunetti-Pierri, N. & Strnad, P. Alpha-1 antitrypsin deficiency: A re-surfacing adult liver disorder. *J Hepatol* **76**, 946-958 (2022).

15 Loomba, R. *et al.* Multi-Society Expert Panel Consensus Guidance Regarding Clinical Assessment and Clinical Trial Endpoints in Adults With Alpha-1 Antitrypsin Deficiency Associated Liver Disease. *Gastroenterology* (2025).

16 Turner, A. M. *et al.* Advancing the understanding and treatment of lung pathologies associated with alpha 1 antitrypsin deficiency. *Ther Adv Respir Dis* **19**, 17534666251318841 (2025).

17 Stockley, R. A. & Parr, D. G. Antitrypsin deficiency: still more to learn about the lung after 60 years. *ERJ Open Research* **10**, 00139-02024 (2024).

18 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).

19 Blanco, I. *et al.* Alpha-1 antitrypsin Pi*Z gene frequency and Pi*ZZ genotype numbers worldwide: an update. *Int J Chron Obstruct Pulmon Dis* **12**, 561-569 (2017).

20 Wiesemann, G. S., Oshins, R. A., Flagg, T. O. & Brantly, M. L. Novel SERPINA1 Alleles Identified through a Large Alpha-1 Antitrypsin Deficiency Screening Program and Review of Known Variants. *Chronic Obstr Pulm Dis* **10**, 7-21 (2023).

21 Wang, C., Zhao, P., Sun, S., Teckman, J. & Balch, W. E. Leveraging Population Genomics for Individualized Correction of the Hallmarks of Alpha-1 Antitrypsin Deficiency. *Chronic Obstr Pulm Dis* **7**, 224-246 (2020).

22 Giacopuzzi, E. *et al.* Real-world clinical applicability of pathogenicity predictors assessed on SERPINA1 mutations in alpha-1-antitrypsin deficiency. *Hum Mutat* **39**, 1203-1213 (2018).

23 Carrell, R. W. & Lomas, D. A. Alpha1-antitrypsin deficiency--a model for conformational diseases. *N Engl J Med* **346**, 45-53 (2002).

24 Bouton, M. C. *et al.* The under-appreciated world of the serpin family of serine proteinase inhibitors. *EMBO Mol Med* **15**, e17144 (2023).

25 D'Acunto, E. *et al.* Neuroserpin: structure, function, physiology and pathology. *Cellular and Molecular Life Sciences* **78**, 6409-6430 (2021).

26 Maurer, M. *et al.* The international WAO/EAACI guideline for the management of hereditary angioedema - The 2021 revision and update. *World Allergy Organ J* **15**, 100627 (2022).

27 Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**, D1062-D1067 (2018).

28 Ottaviani, S. *et al.* Comprehensive Clinical Diagnostic Pipelines Reveal New Variants in Alpha-1 Antitrypsin Deficiency. *Am J Respir Cell Mol Biol* **69**, 355-366 (2023).

29 Fernandez-Escamilla, A. M., Rousseau, F., Schymkowitz, J. & Serrano, L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol* **22**, 1302-1306 (2004).

30 Zambrano, R. *et al.* AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures. *Nucleic Acids Research* **43**, W306-W313 (2015).

31 Navarro, S. & Ventura, S. Computational methods to predict protein aggregation. *Current Opinion in Structural Biology* **73**, 102343 (2022).

32 Ke, P. C. *et al.* Half a century of amyloids: past, present and future. *Chem Soc Rev* **49**, 5473-5509 (2020).

33 Faull, S. V. *et al.* The structural basis for Z α(1)-antitrypsin polymerization in the liver. *Sci Adv* **6** (2020).

34 Vickers, S. *et al.* Top-Down Ion Mobility Mass Spectrometry Reveals a Disease Associated Conformational Ensemble of Alpha-1-antitrypsin. *J Am Chem Soc* **147**, 16909-16921 (2025).

35 Brzozowska, N. *et al.* Selection for somatic escape variants in SERPINA1 in the liver of patients with alpha-1 antitrypsin deficiency. *Nature Genetics* **57**, 875-883 (2025).

36 Sun, S. *et al.* Capturing the conversion of the pathogenic alpha-1-antitrypsin fold by ATF6 enhanced proteostasis. *Cell Chem Biol* **30**, 22-42.e25 (2023).

37 Zhao, P., Wang, C., Sun, S., Wang, X. & Balch, W. E. Tracing genetic diversity captures the molecular basis of misfolding disease. *Nat Commun* **15**, 3333 (2024).

38 Sun, S. *et al.* Spatial covariance reveals isothiocyanate natural products adjust redox stress to restore function in alpha-1-antitrypsin deficiency. *Cell Rep Med* **6**, 101917 (2025).

39 Miravitlles, M., Anzueto, A. & Barrecheguren, M. Nine controversial questions about augmentation therapy for alpha-1 antitrypsin deficiency: a viewpoint. *European Respiratory Review* **32**, 230170 (2023).

40 Brantly, M. Treatment for Alpha-1 Antitrypsin Deficiency: Does Augmentation Therapy Work? *Am J Respir Crit Care Med* **208**, 948-949 (2023).

41 Lomas, D. A. *et al.* Development of a small molecule that corrects misfolding and increases secretion of Z α(1) -antitrypsin. *EMBO Mol Med* **13**, e13167 (2021).

720 42 Ronzoni, R. *et al.* Susceptibility of alpha-1 antitrypsin deficiency variants to
721 polymer-blocking therapy. *JCI Insight* **10** (2025).
722 43 Balch, W. E., Morimoto, R. I., Dillin, A. & Kelly, J. W. Adapting proteostasis for
723 disease intervention. *Science* **319**, 916-919 (2008).
724 44 Karatas, E. & Bouchecareilh, M. Alpha 1-Antitrypsin Deficiency: A Disorder of
725 Proteostasis-Mediated Protein Folding and Trafficking Pathways. *Int J Mol Sci* **21** (2020).
726 45 Wiseman, R. L., Mesgarzadeh, J. S. & Hendershot, L. M. Reshaping endoplasmic
727 reticulum quality control through the unfolded protein response. *Mol Cell* **82**, 1477-1491
728 (2022).
729 46 Khodayari, N. *et al.* Modulation of calreticulin expression reveals a novel exosome-
730 mediated mechanism of Z variant α(1)-antitrypsin disposal. *J Biol Chem* **294**, 6240-6252
731 (2019).
732 47 Khodayari, N. *et al.* Erdj3 Has an Essential Role for Z Variant Alpha-1-Antitrypsin
733 Degradation. *J Cell Biochem* **118**, 3090-3101 (2017).
734 48 Prabakaran, R., Rawat, P., Kumar, S. & Michael Gromiha, M. ANuPP: A Versatile
735 Tool to Predict Aggregation Nucleating Regions in Peptides and Proteins. *J Mol Biol* **433**,
736 166707 (2021).
737 49 Kulshreshtha, S., Chaudhary, V., Goswami, G. K. & Mathur, N. Computational
738 approaches for predicting mutant protein stability. *Journal of Computer-Aided Molecular*
739 *Design* **30**, 401-412 (2016).
740 50 Delgado, J., Radusky, L. G., Cianferoni, D. & Serrano, L. FoldX 5.0: working with
741 RNA, small molecules and a new graphical interface. *Bioinformatics* **35**, 4168-4169 (2019).

742 51 Dieckhaus, H., Brocidiacono, M., Randolph, N. Z. & Kuhlman, B. Transfer
743 learning to leverage larger datasets for improved prediction of protein stability changes.
744 *Proceedings of the National Academy of Sciences* **121**, e2314853121 (2024).
745 52 Wang, C., Anglès, F. & Balch, W. E. Triangulating variation in the population to
746 define mechanisms for precision management of genetic disease. *Structure* **30**, 1190-
747 1207.e1195 (2022).
748 53 Hie, B., Bryson, B. D. & Berger, B. Leveraging Uncertainty in Machine Learning
749 Accelerates Biological Discovery and Design. *Cell Syst* **11**, 461-477.e469 (2020).
750 54 Wang, C. & Balch, W. E. Bridging Genomics to Phenomics at Atomic Resolution
751 through Variation Spatial Profiling. *Cell Rep* **24**, 2013-2028.e2016 (2018).
752 55 Romero, P. A., Krause, A. & Arnold, F. H. Navigating the protein fitness landscape
753 with Gaussian processes. *Proc Natl Acad Sci U S A* **110**, E193-201 (2013).
754 56 Patschull, A. O. *et al.* Therapeutic target-site variability in alpha1-antitrypsin
755 characterized at high resolution. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **67**,
756 1492-1497 (2011).
757 57 Yamasaki, M., Sendall, T. J., Pearce, M. C., Whisstock, J. C. & Huntington, J. A.
758 Molecular basis of α1-antitrypsin deficiency revealed by the structure of a domain-
759 swapped trimer. *The EMBO Reports* **12**, 1011-1017 (2011).
760 58 Matamala, N. *et al.* Characterization of Novel Missense Variants of SERPINA1
761 Gene Causing Alpha-1 Antitrypsin Deficiency. *Am J Respir Cell Mol Biol* **58**, 706-716
762 (2018).

763 59 Kueppers, F. *et al.* Protein modeling to assess the pathogenicity of rare variants of
764 SERPINA1 in patients suspected of having Alpha 1 Antitrypsin Deficiency. *BMC Med*
765 *Genet* **20**, 125 (2019).
766 60 Carpagnano, G. E. *et al.* A New SERPINA-1 Missense Mutation Associated with
767 Alpha-1 Antitrypsin Deficiency and Bronchiectasis. *Lung* **195**, 679-682 (2017).
768 61 Kueppers, F. Clinical presentations of four patients with rare Alpha 1 Antitrypsin
769 variants identified in a single US center. *Respir Med Case Rep* **32**, 101345 (2021).
770 62 Olivares Rivera, A. *et al.* Unveiling Genetic Diversity in Alpha-1 Antitrypsin
771 Deficiency: Case Report of Two Siblings with the Novel SERPINA1 Mutation
772 Pi*c.236T>A. *European Respiratory Journal* **66**, PA345 (2025).
773 63 Oliveira, M. J. *et al.* Alpha-1 antitrypsin deficiency caused by a novel mutation
774 (p.Leu263Pro): Pi*ZQ0gaia - Q0gaia allele. *Rev Port Pneumol (2006)* (2015).
775 64 Haq, I. *et al.* Deficiency Mutations of Alpha-1 Antitrypsin. Effects on Folding,
776 Function, and Polymerization. *Am J Respir Cell Mol Biol* **54**, 71-80 (2016).
777 65 Gonzalez, A. *et al.* New variants of alpha-1-antitrypsin: structural simulations and
778 clinical expression. *Respir Res* **23**, 339 (2022).
779 66 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold.
780 *Nature* **596**, 583-589 (2021).
781 67 Wang, C. & Balch, W. E. in *Alpha-1 Antitrypsin: Role in Health and Disease* (eds
782 Adam Wanner & Robert A. Sandhaus) 53-83 (Springer International Publishing, 2016).
783 68 Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with
784 AlphaFold 3. *Nature* **630**, 493-500 (2024).

785 69 Wang, C., Elghobashi-Meinhardt, N. & Balch, W. E. Covariant Fitness Clusters
786 Reveal Structural Evolution of SARS-CoV-2 Polymerase Across the Human Population.
787 *bioRxiv, 2022.01.07.475295* (2022).

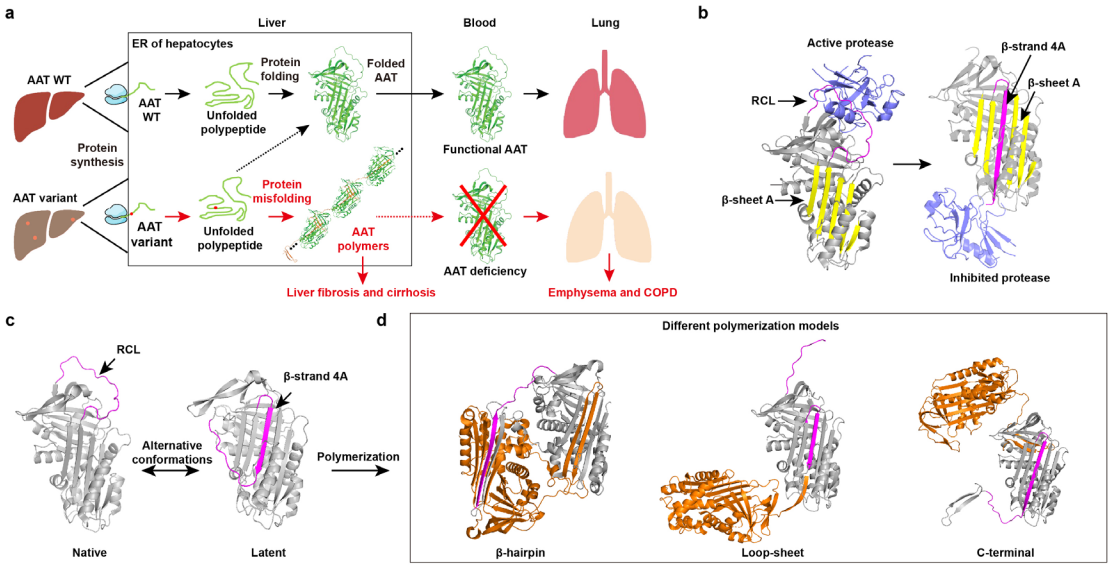


Figure 1. AAT polymerization and AAT deficiency. a, Physiological folding and trafficking of wild-type AAT (WT) versus the consequences of polymerogenic variants. AAT-WT folds in the endoplasmic reticulum (ER) of hepatocytes, is secreted into the bloodstream and protects the lung from neutrophil elastase-mediated damage. Polymerogenic variants misfold and are retained in the ER, accumulating as intracellular polymers that contribute to liver fibrosis and cirrhosis, while reduced secretion of functional AAT leads to emphysema and chronic obstructive pulmonary disease (COPD). **b**, Serpin protease inhibition mechanism. Cleavage of the reactive center loop (RCL) by a target protease triggers insertion of the cleaved loop as β-strand 4A into β-sheet A and translocation of the protease to form an inhibited complex. **c**, Native and latent AAT conformations. Transition to the latent state involves insertion of the uncleaved RCL into β-sheet A. **d,** Representative models proposed for AAT polymerization, including β-hairpin, loop–sheet and C-terminal domain-swap mechanisms.

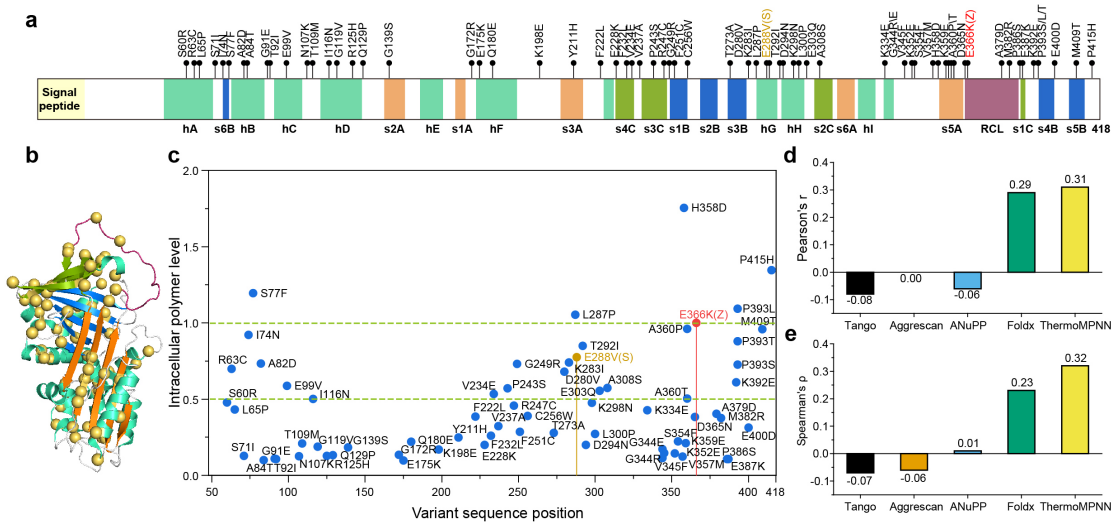
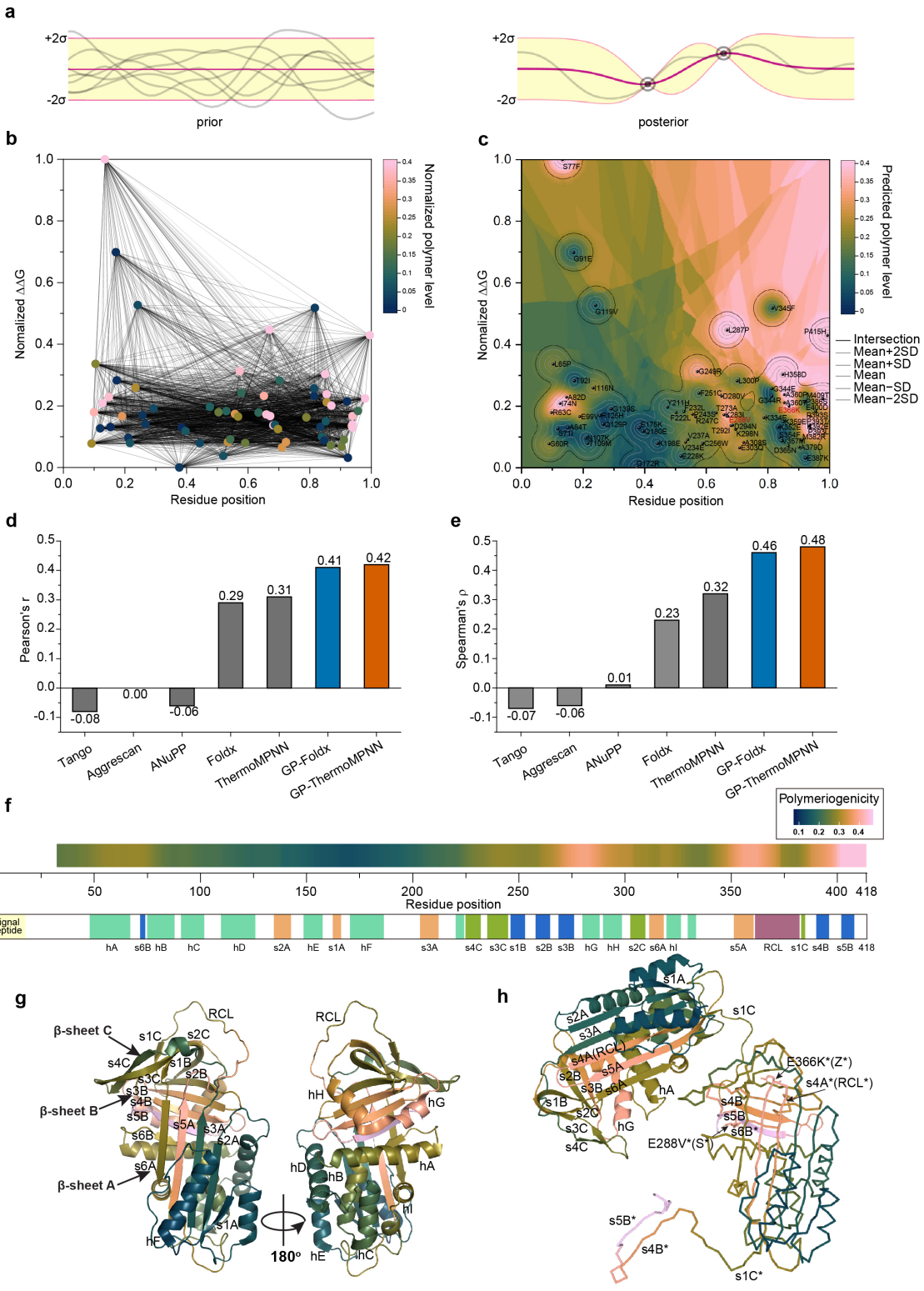


Figure 2. Existing aggregation predictors underperform on AAT variants. a, Distribution of the 68 characterized AAT missense variants along the AAT sequence. Secondary structure elements are annotated, with α-helices (*hA-hI*) shown as cyan blocks and β-sheets A (*s1-6A*), B (*s1-6B*) and C (*s1-4C*) shown in orange, blue and green, respectively. The reactive center loop (*RCL*) is shown in purple. **b**, Locations of these variants mapped onto the AAT structure (PDB 3NE4). Variant residues are indicated by yellow spheres (Cα atoms), with α-helices and β-sheets A, B and C highlighted. **c,** Previously measured intracellular polymer burden for the 68 missense AAT variants plotted against sequence positions. Polymer levels are normalized to AAT-Z. AAT-Z and AAT-S variants are highlighted with red and yellow labels, respectively. **d, e**, Pearson (**d**) and Spearman (**e**) correlations between experimentally measured polymer burden and predicted scores from TANGO, AggreScan3D, ANuPP, FoldX, and ThermoMPNN across the 68 AAT variants.

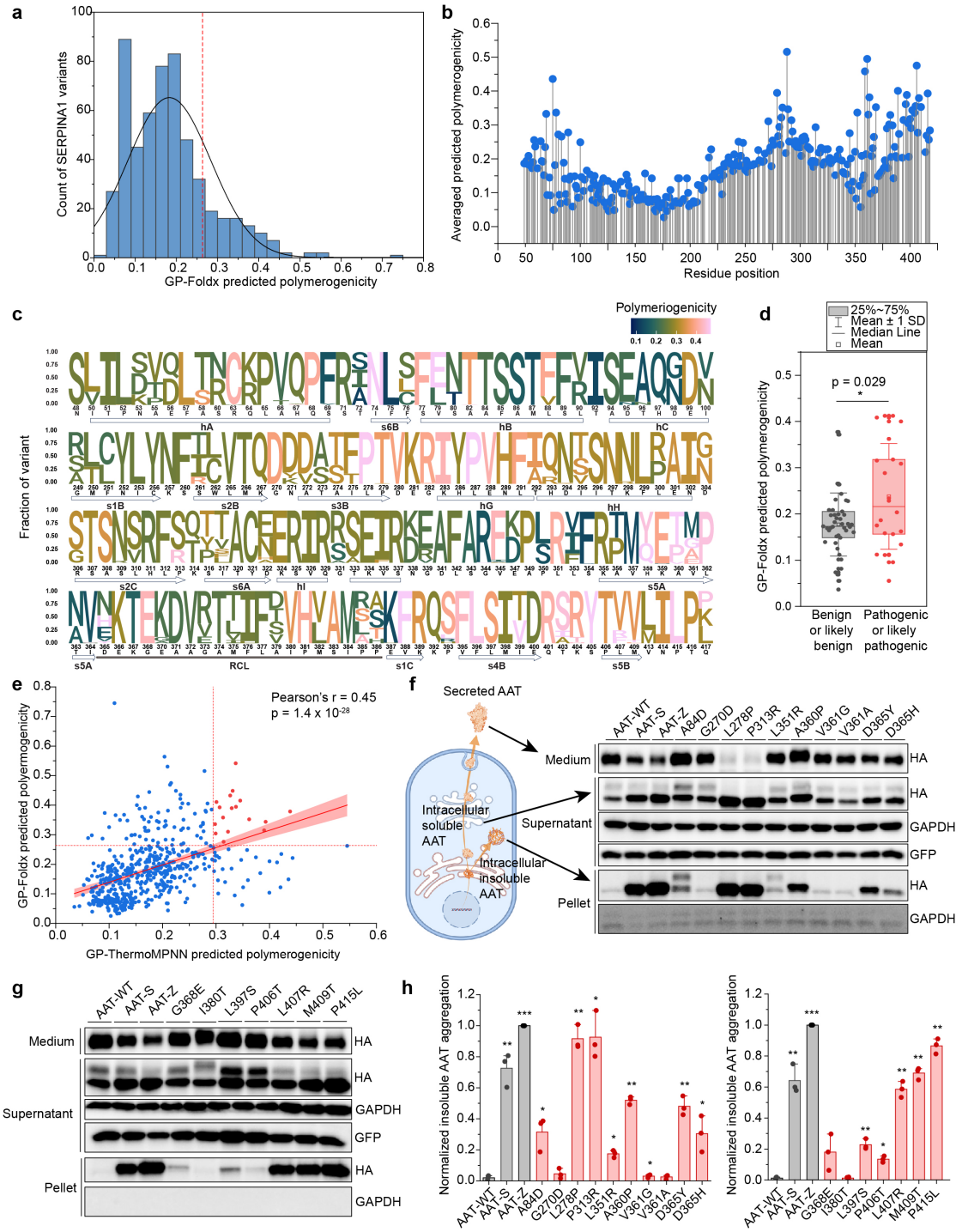


817

818 **Figure 3. Gaussian process mapping of the stability-dependent AAT polymerization**

819 **landscape. a,** Schematic of GP prior and posterior distributions before and after

conditioning on observations. **b,** Input feature space for AAT variants, with normalized residue position (*x* axis), stability perturbation ΔΔG (*y* axis), and experimentally measured polymer burden (*z* axis, color scale). Black lines indicate pairwise distances between variants in the joint (*x, y*) space. **c**, GP-inferred AAT polymerization landscape relating ΔΔG (*y*-axis) to polymer burden (color scale) across all residues of the AAT polypeptide (*x*-axis). Contour lines denote the mean and standard deviation (SD) of the low-variance component identified by a two-component Gaussian mixture model fitted to the GP posterior variance, highlighting higher-confidence regions. **d, e,** Pearson (**d**) and Spearman (**e**) correlations of the leave-one-out cross-validation results of GP-FoldX and GP-ThermoMPNN models compared with other models. **f,** Residue-level polymerogenicity scores obtained by inverse variance weighting of GP posterior means across queried points assigned to each residue position. **g**, **h**, Mapping the residue-level polymerogenicity scores to the AAT monomer structure (PDB: 3NE4) (**g**) and polymer structure (PDB: 3T1P) (**h**).



833

834 **Figure 4. Population-scale prediction and experimental validation of polymerogenic**

835 **AAT variants. a,** Distribution of GP-FoldX predicted polymer burden for AAT missense

836 variants reported in gnomAD. The GP leave-one-out posterior mean for the Z allele is

837 indicated by the red dashed line. **b,** Residue-wise average of predicted polymer burden for
838 gnomAD AAT variants across sequence positions. **c,** Residue-level population variant
839 signatures and their predicted polymer burden. Residues 48-100 and 249-417 are shown
840 with secondary structure annotations. **d,** Predicted polymer burden for ClinVar-annotated
841 benign/likely benign versus pathogenic/likely pathogenic AAT variants. P values were
842 computed using a two-tailed Welch's t-test, *p<0.05. **e,** Correlation between the GP-FoldX
843 and GP-ThermoMPNN predictions. Variants with predicted values exceeding the leave-
844 one-out Z-allele posterior mean (red dashed lines) were selected for experimental testing
845 (red). **f-g** Experimental validation of predicted polymerogenic AAT variants in Huh7 cells.
846 HA-tagged AAT (WT and variants) was transiently expressed and analyzed by
847 immunoblotting in conditioned media, soluble lysate (supernatant) and insoluble fraction
848 (pellet). GFP (IRES–GFP) served as a transfection/expression control, and GAPDH was
849 used as a loading control for the soluble fraction and to assess fractionation quality. **h,**
850 Quantification of insoluble AAT polymer signal normalized to the Z variant. For each
851 variant, significance versus WT was assessed using a one-sided paired t-test across three
852 biological replicates with Benjamini–Hochberg false discovery rate (BH–FDR) correction;
853 adjusted q values are shown. *q<0.05, **q<0.01, ***q<0.001.

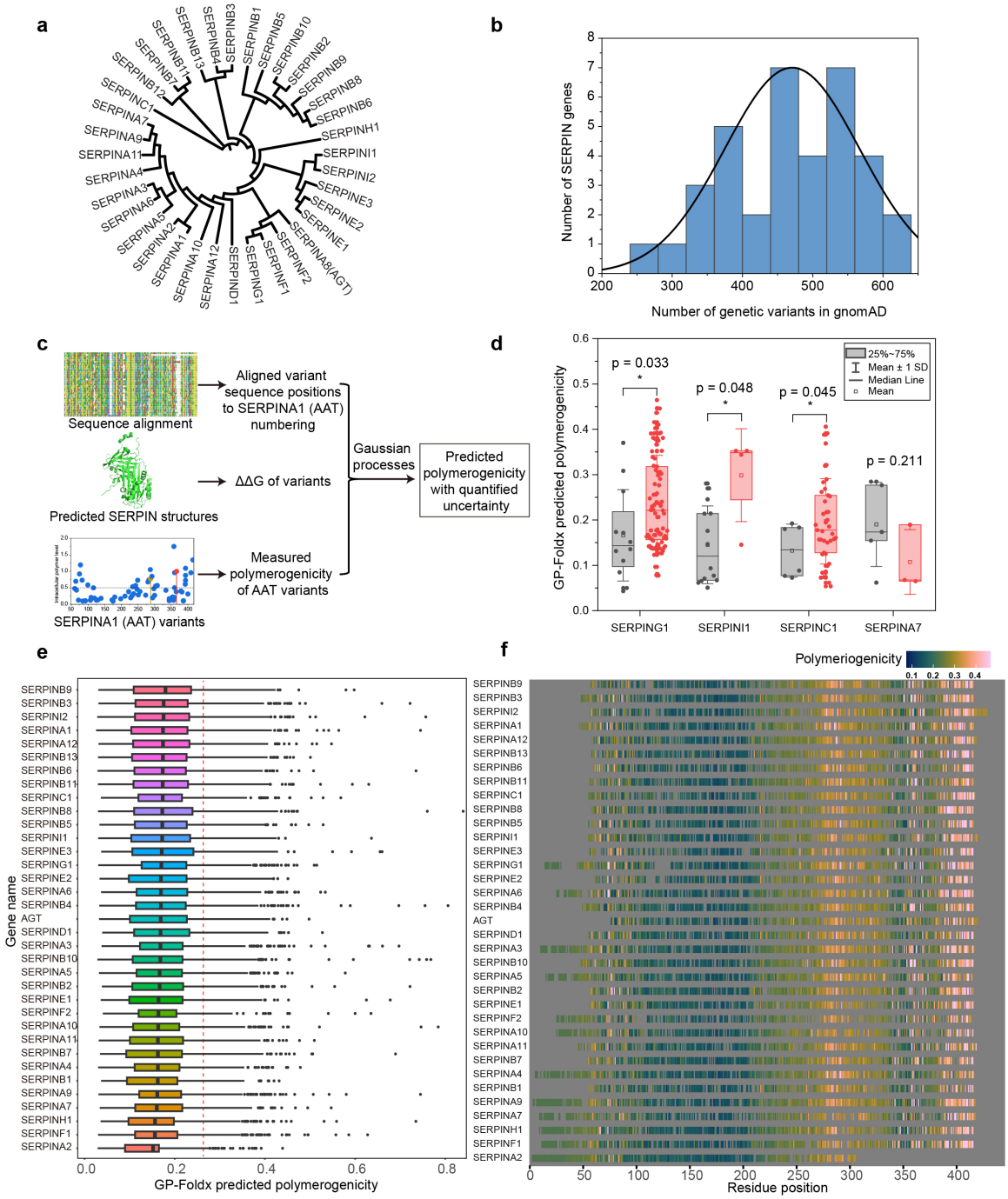
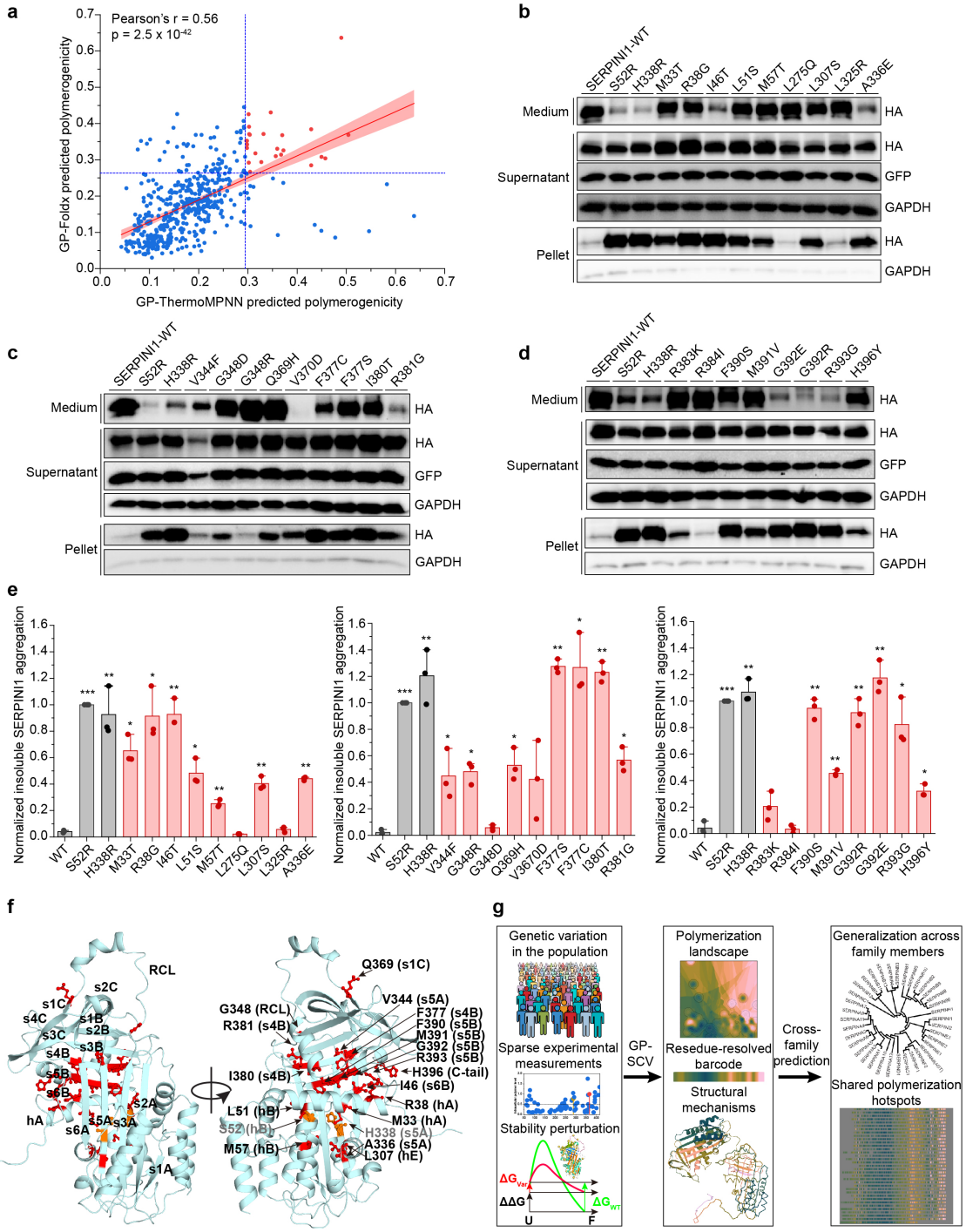


Figure 5. Family-wide prediction of polymerogenicity across human serpin variants. **a,** Phylogenetic relationships among human serpin family members. **b,** Number of missense variants reported in gnomAD for each serpin gene. **c,** Schematic of the family-wide prediction framework. The AAT-trained GP model was applied to other serpins by

mapping residue position to the AAT reference via sequence alignment, computing variant-specific ΔΔG on the query protein, and using the same trained GP without adding non-AAT variants to the training set. **d,** Predicted polymer burden for ClinVar-annotated variants across serpin genes. Variants annotated as benign/likely benign or pathogenic/likely pathogenic are shown in grey and red, respectively. P values were computed using a two-tailed Welch's t-test, *$p<0.05$. **e,** Distribution of predicted polymer burden for gnomAD missense variants in each serpin gene. Boxes indicate the interquartile range (IQR) with the median shown as a horizontal line; whiskers extend to 1.5×IQR, and points denote outliers. The red dashed line marks the leave-one-out posterior mean prediction for the AAT Z allele, used as a reference threshold for elevated polymerization risk. **f,** Alignment-based comparison of predicted residue-level polymerization landscapes across serpins, shown as position-aligned polymer burden profiles.



871

872 **Figure 6. Validation of predicted polymerogenicity of neuroserpin variants. a,**

873 Comparison of polymerogenicity predictions for neuroserpin variants from the GP-FoldX

874 and GP–ThermoMPNN models. Variants with predicted values exceeding the AAT Z-allele

leave-one-out posterior mean (blue dashed lines; used as a reference threshold) were selected for experimental testing (red). **b-d,** HA-tagged neuroserpin (WT and variants) was transiently expressed in HeLa cells and analyzed by immunoblotting in conditioned media, soluble lysate (supernatant) and insoluble fraction (pellet). GFP (IRES–GFP) served as a transfection/expression control, and GAPDH was used as a loading control for the soluble fraction and to assess fractionation quality. WT neuroserpin serves as negative control while FENIB variants S52R and H338R are shown as positive controls for insoluble polymer formation. **e,** Quantification of insoluble neuroserpin polymer signal normalized to the S52R variant. For each variant, significance versus WT was assessed using a one-sided paired t-test across three biological replicates with Benjamini–Hochberg false discovery rate (BH–FDR) correction; adjusted q values are shown. *q<0.05, **q<0.01, ***q<0.001. **f,** Variants with significantly higher insoluble neuroserpin signal than WT are mapped onto the AlphaFold2-predicted neuroserpin structure as red sticks and balls. Known FENIB variants S52R and H338R are highlighted in orange. **g,** GP–SCV framework for family-wide serpin polymerization prediction. Sparse experimental measurements, population variation and stability perturbations (ΔΔG) are integrated with a GP-SCV model to infer a residue-resolved polymerization landscape and structural mechanisms of polymerization for AAT. The learned landscape is then transferred across serpin family members to predict shared polymerization hotspots.

Supplementary information:

Mapping the polymerization landscape of human serpins across genetic variation using Gaussian Processes

Authors: Shuyuan Chen[1,2], Yunsong Deng[2,3,4], Xin Zhang[1#] and Chao Wang[2,3#*]

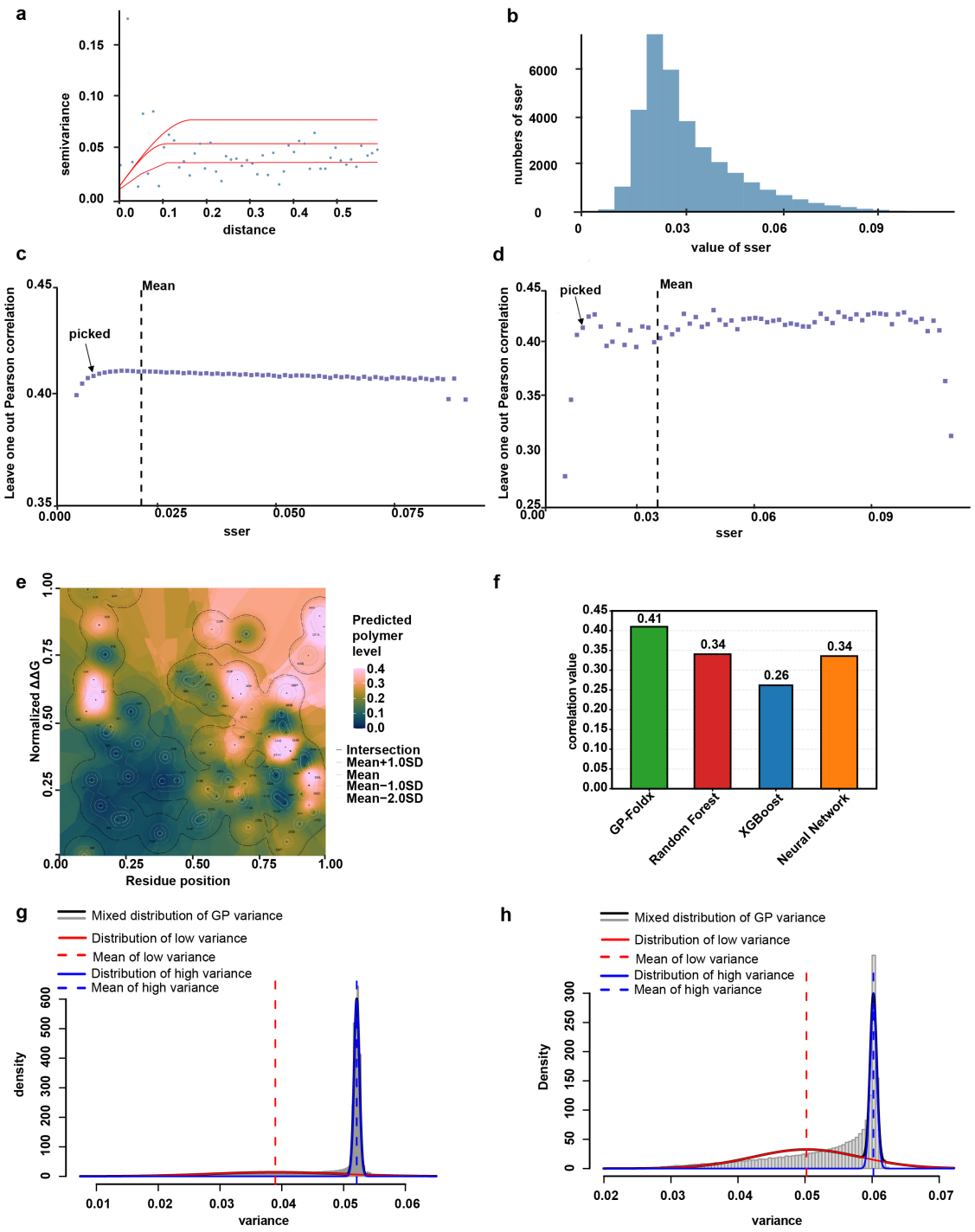
Affiliations: [1]Department of Chemistry, Westlake University, Hangzhou 310030, China. [2]Institute of Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen, 518132, China. [3]Shenzhen Medical Academy of Research and Translation (SMART), Shenzhen 518107, China. [4]Westlake University, Hangzhou 310030, China.

[#]Correspondence: chaowang@szbl.ac.cn; zhangxin@westlake.edu.cn

[*]Lead contact: chaowang@szbl.ac.cn

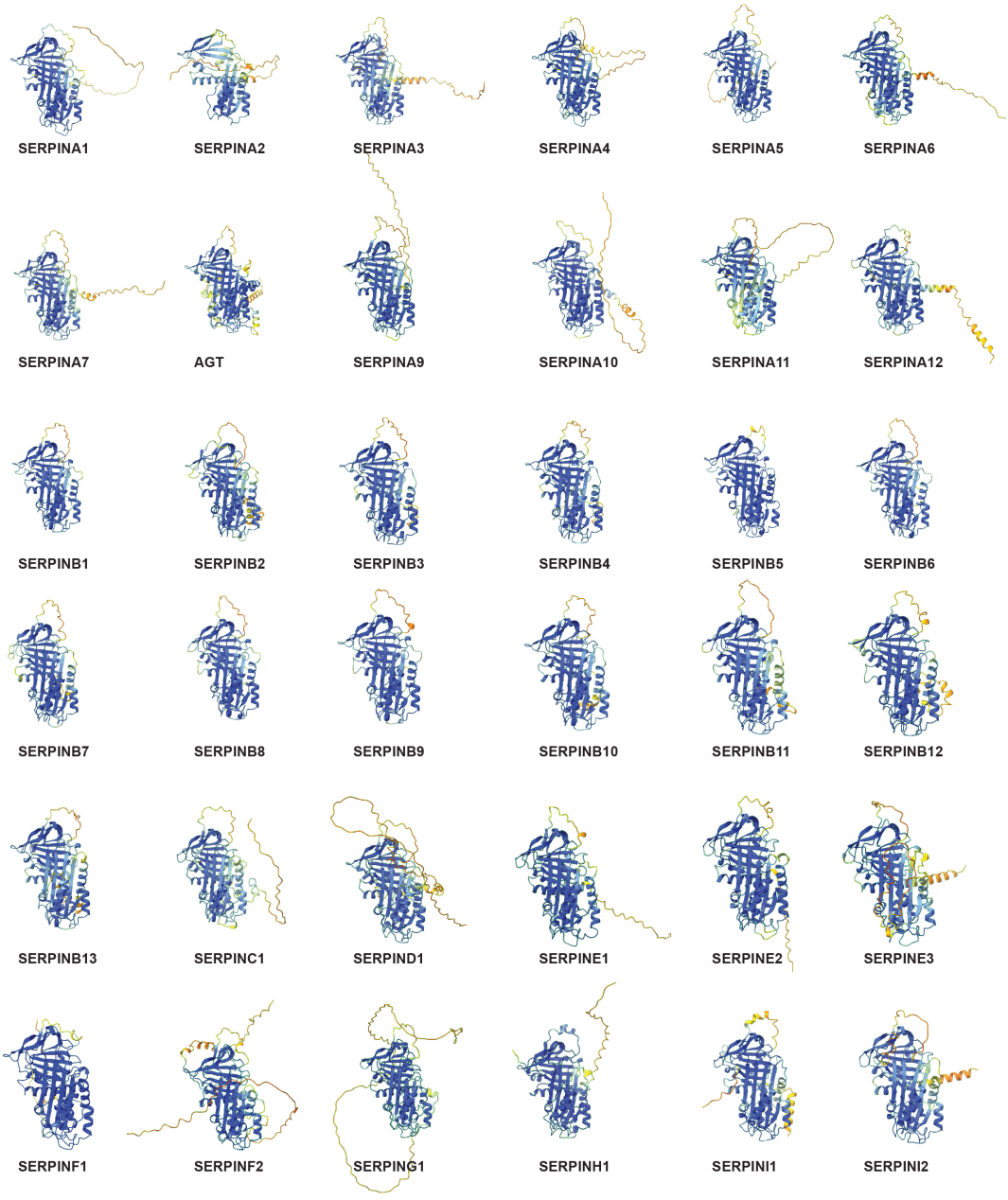
The supplementary information includes:

Supplementary Figure 1 and 2



Supplementary Fig. 1. Hyperparameter selection, model benchmarking and uncertainty characterization. a, Representative variogram fits illustrating models with high, intermediate and low variogram fit error, quantified as the sum of squared errors (sser)

between the empirical (sample) variogram and the fitted variogram model. **b,** Distribution of sser values across all candidate models evaluated during hyperparameter search. **c,d,** Relationship between variogram fit (sser) and leave-one-out cross-validation (LOOCV) performance for GP-FoldX (**c**) and GP-ThermoMPNN (**d**). Candidate models were binned by sser (bin width = 0.1 standard deviations of sser), and the highest LOOCV Pearson correlation within each bin is shown. The mean sser is indicated by a black dashed line. The selected model corresponds to the onset of the LOOCV plateau at low sser (the "elbow" point indicated by arrows), prioritizing the smallest sser among models with near-saturated LOOCV Pearson correlation. **e,** The stability-dependent AAT polymerization landscape inferred by the GP-ThermoMPNN model, showing predicted polymer level (*z*-axis, color scale) as a function of normalized residue position (*x*-axis) and normalized ΔΔG (*y*-axis). Contours indicate the Gaussian-mixture–based uncertainty threshold (intersection) and offsets around the low-variance component (mean and ± s.d.), highlighting higher-confidence regions. **f,** LOOCV Pearson correlation for GP-FoldX compared with alternative regressors (random forest, XGBoost and neural network) trained on the same inputs. **g-h,** Distributions of GP posterior variance for the GP-FoldX (**g**) and GP-ThermoMPNN (**h**) models across queried landscape points, modeled with a two-component Gaussian mixture. The fitted mixture (black line on grey histograms) and the low-variance (red) and high-variance (blue) components are shown, with dashed vertical lines indicating the component means.



Supplementary Fig. 2. AlphaFold2-predicted structures of human serpin family members. AlphaFold2 models for each serpin were obtained from the AlphaFold Protein Structure Database.