Machine Learning–Based Prediction of Transthyretin Binding Activity and its Application to Screening Food-derived Compounds for ATTR Amyloidosis

Yuma Iwashita and Yoshihiro Uesawa *

Posted Date: 16 April 2026

doi: 10.20944/preprints202604.1199.v1

Keywords: transthyretin amyloidosis; wild-type ATTR; transthyretin stabilizer; machine learning; in silico screening; food-derived compounds; polyphenols

 doi:10.20944/preprints202604.1199.v1

Article

Machine Learning–Based Prediction of Transthyretin Binding Activity and Its Application to Screening Food-Derived Compounds for ATTR Amyloidosis

Yuma Iwashita and Yoshihiro Uesawa *

Department of Medical Molecular Informatics, Meiji Pharmaceutical University, Japan; e-mail@e-mail.com

* Correspondence: uesawa@my-pharm.ac.jp; Tel.: (optional; include country code)

Abstract

Transthyretin amyloidosis (ATTR) is a progressive disease caused by the dissociation of the transthyretin (TTR) tetramer, leading to amyloid fibril formation. Although pharmacological stabilizers have been developed, preventive strategies for wild-type ATTR (ATTRwt) have not been established. This study developed a computational model to predict TTR binding activity from chemical structure and to apply the model to screen food-derived compounds as potential preventive candidates. A machine learning model was constructed using TTR-8-anilino-1-naphthalenesulfonic acid displacement assay data from the Tox24 Challenge. The model achieved root mean square error and R^2 values of 21.34 and 0.64, respectively, on an external test dataset. Using an integrated dataset compiled from multiple literature sources, the predicted TTR binding activity exhibited a significant positive correlation with amyloid fibril formation inhibition (Spearman's $\rho = 0.602$, $p < 0.001$). The model was then applied to the PhytoHub database, identifying 63 candidate compounds with high predicted binding activity, predominantly polyphenols, found in 126 food sources. These results suggest that the proposed in silico method is useful for identifying potential TTR stabilizers from food-derived compounds and may contribute to the exploration of effective preventive strategies for ATTRwt.

Keywords: transthyretin amyloidosis; wild-type ATTR; transthyretin stabilizer; machine learning; in silico screening; food-derived compounds; polyphenols

1. Introduction

Transthyretin amyloidosis (ATTR) is a form of amyloidosis caused by transthyretin (TTR), broadly classified into the hereditary (ATTRv) and wild-type (ATTRwt) forms [1]. In ATTRv, clinical manifestations, e.g., familial amyloidotic polyneuropathy and familial amyloid cardiomyopathy, are well recognized [2,3]. In contrast, ATTRwt was formerly referred to as senile systemic amyloidosis, and its clinical features have been described accordingly [4].

Amyloid formation from TTR proceeds via an initial dissociation of the tetrameric structure [5], and this tetramer dissociation represents the rate-limiting step in pathogenesis, after which the released monomers undergo partial structural denaturation, becoming amyloidogenic [6]. Ultimately, these monomers aggregate to form amyloid fibrils and amorphous deposits [7].

In ATTRwt, amyloid deposition in the myocardium leads to impaired diastolic function and progressive congestive heart failure. Clinically, the disease predominantly manifests as diastolic dysfunction, frequently presenting as heart failure with preserved ejection fraction. In addition, carpal tunnel syndrome and spinal stenosis may precede cardiac symptoms [8].

Based on this pathogenic mechanism, pharmacological agents that stabilize TTR and inhibit the progression of amyloid formation have been approved. A representative drug is tafamidis, which is indicated for ATTRwt cardiomyopathy, ATTRv cardiomyopathy, and early-stage ATTRv polyneuropathy [1]. However, the treatment cost is substantial. For example, in the United States, the annual cost of tafamidis has been reported to reach approximately $225,000 [9]. Thus, long-term

 doi:10.20944/preprints202604.1199.v1

continuous therapy can impose significant economic burdens on patients and healthcare insurance systems. Furthermore, no established preventive strategies currently exist for ATTRwt, and knowledge about effective interventions at the presymptomatic stage remains limited. Thus, investigating potential preventive interventions that may contribute to reducing the risk of onset or delay the progression of ATTRwt is important.

Research on dietary interventions designed to prevent ATTRwt is limited. However, a previous study demonstrated that arginine stabilizes the TTR tetramer and that oral supplementation of 5000 mg per day for 5 days increased the TTR tetramer/monomer ratio, suggesting the potential of supplementation as a preventive intervention [10]. That study suggested that the TTR stabilization mechanism by arginine may involve interactions with aromatic amino acid residues that are abundant in TTR, particularly tryptophan.

Furthermore, although specifically targeted at disease progression inhibition (rather than ATTRwt prevention), epigallocatechin-3-gallate (EGCG), which is a component of green tea, has been reported to stabilize TTR [11]. A clinical study demonstrated that daily consumption of 1.5–2 L of green tea over a 12-month period resulted in an approximately 13% reduction in left ventricular myocardial mass [12]. TTR stabilization by EGCG is attributed to tetramer stabilization via binding at the TTR dimer–dimer interface, which indicates that food-derived compounds can potentially stabilize TTR [11].

Generally, the degree of pharmacological kinetic stabilization of TTR depends on the fraction of TTR tetramers in which at least one binding site is occupied by a stabilizer [13]. This fraction is determined by the binding affinity of the stabilizer for TTR, the concentration of albumin, i.e., the major plasma protein competing for binding with TTR, and the plasma concentrations of the stabilizer, TTR, and albumin.

Among these factors, the binding affinity for TTR is considered an important determinant reflecting TTR stabilization capacity. Thus, constructing a model to predict an index related to TTR binding affinity may enable the effective estimation of an index reflecting TTR stabilization capacity. TTR-ANSA displacement activity, based on the displacement of the fluorescent probe 8-anilino-1-naphthalenesulfonic acid (ANSA), has been reported as an indicator of TTR binding affinity [14].

The exploration of natural compounds with TTR-binding properties has been previously reported [15]. However, these compounds have primarily been studied for their inhibitory activity against TTR amyloid formation rather than for preventive purposes as food-derived compounds. Food-derived compounds are consumed daily, have high safety profiles, and offer the advantage of long-term intake. ATTRwt is an age-related disease; thus, even compounds with weaker stabilization effects than pharmaceutical agents may contribute to disease prevention through cumulative long-term intake [16]. From this perspective, the evaluation of food-derived compounds is important. Therefore, developing a computational screening method based on chemical structural information to efficiently identify candidate compounds with TTR stabilization activity in food-derived compounds may be useful.

In this study, we used TTR-ANSA displacement activity as an indicator of TTR binding activity and attempted to construct a model to predict this activity from chemical structural information.

Hypothesis 1: TTR binding activity (%) can be predicted from chemical structural information.

Hypothesis 2: The TTR binding activity score predicted from the chemical structure exhibits a monotonic association with the amyloid fibril formation inhibition rate reported in previous studies under acidic conditions.

By testing these hypotheses, this study aimed to evaluate the utility of a computational screening approach to efficiently identify candidate compounds with TTR stabilization activity from food-derived compounds. In addition, the proposed method was applied to a food compound database for exploratory screening. Figure 1 shows an overview of the analysis.

 doi:10.20944/preprints202604.1199.v1

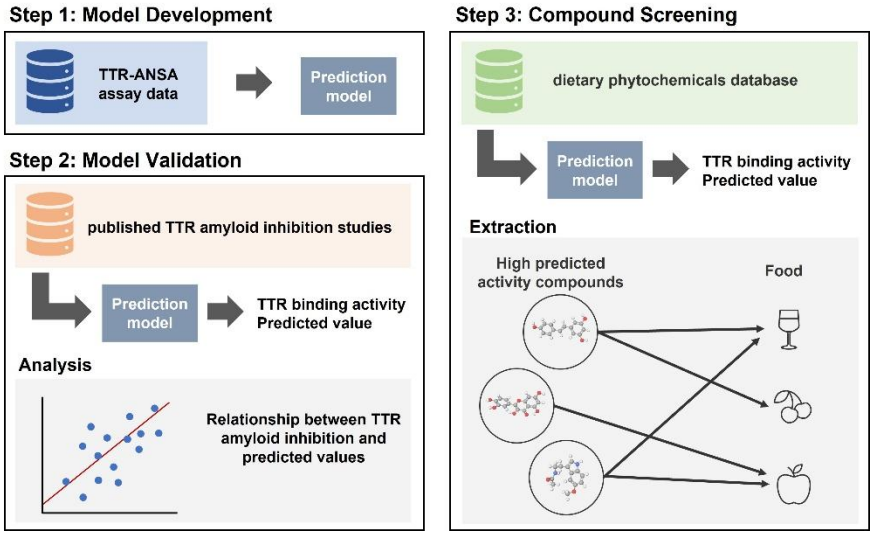

Figure 1. Overview of analytical workflow. Step 1: A TTR binding activity prediction model was constructed using TTR-ANSA assay data. Step 2: Model validity was verified using TTR amyloid formation inhibition data collected from the literature, and the relationship between the predicted values and inhibitory activity was analyzed. Step 3: The constructed model was applied to a food-derived compound database to extract compounds with high predicted activity and identify foods containing these compounds.

2. Results

2.1. Construction of TTR Binding Activity Prediction Model

To construct the TTR binding activity prediction model, TTR-ANSA displacement activity data published in the 2024 Tox24 Challenge were obtained from the Online Chemical Modeling Environment (OCHEM; https://ochem.eu/static/challenge-data.do on May 29, 2025) [17]. In this study, the train (N = 1012) and leaderboard (N = 200) datasets were employed as the training dataset, and the blind (N = 300) dataset was employed as the test dataset for external validation.

The TTR binding activity was measured by the ANSA displacement assay as relative activity (%) with respect to thyroxine (T4) [14]. In addition, the range of the TTR binding activity in the training dataset was −45.0 to 111.1% with a mean of 39.6% and a standard deviation (SD) of 36.1% (Table S1).

2.1.1. Model Selection

To evaluate the effect of molecular descriptor input formats, differing in whether salt/complex processing was applied, on model performance, three input formats were compared, i.e., (a) molecular descriptors calculated from structures without salt/complex processing, (b) molecular descriptors calculated from structures after salt/complex processing, and (c) an integrated set combining molecular descriptors from preprocessing and postprocessing structures. All molecular descriptors were calculated using the Mordred descriptor package [18]. The predictive performance of these input formats was then evaluated using LightGBM, which is an implementation of gradient boosting [19].

Model construction was performed in two stages, i.e., feature selection and hyperparameter optimization. Following feature preprocessing and selection, the final numbers of molecular descriptors adopted for formats (a), (b), and (c) were 443, 522, and 820, respectively (Table 1).

 doi:10.20944/preprints202604.1199.v1

Table 1. Progression of descriptor counts through preprocessing stages, including missing value removal and zero-variance filtering, followed by feature selection based on feature importance.

Input format	Number of descriptors					
	Total	Preprocessing			Feature selection	
		Missing values	Zero variance	After filtering	Importance = 0	Final
(a)	1613	884	119	610	167	444
(b)	1613	729	123	761	239	522
(c)	3226	1613	242	1371	551	820

Hyperparameter optimization was then performed using these descriptors to determine the optimal parameter set. Here, the mean root mean square error (RMSE) in five-fold cross-validation (CV) for the training dataset was 23.42, 23.25, and 22.82 for formats (a), (b), and (c), respectively, with format (c) obtaining the highest predictive accuracy (Table 2).

Table 2. Comparison of predictive accuracy (RMSE) across input formats in five-fold CV.

Input format	RMSE						
	fold_1	fold_2	fold_3	fold_4	fold_5	Mean	Std
(a)	22.53	23.17	24.79	23.82	22.81	23.42	0.90
(b)	22.37	24.00	23.04	23.59	23.23	23.25	0.61
(c)	22.17	22.39	24.34	22.89	22.33	22.82	0.89

Thus, format (c) was adopted as the final model, and predictions were performed on the test dataset, which was not used during the training phase, yielding an RMSE value of 21.34. Here, the R^2 value, which was utilized as a reference metric, was 0.64 (Figure 2). In the Tox24 Challenge, models that obtained an RMSE value of 21.4 or less on the test dataset were reported as the top-performing group, and the performance obtained in this study was comparable to these reported results [20].

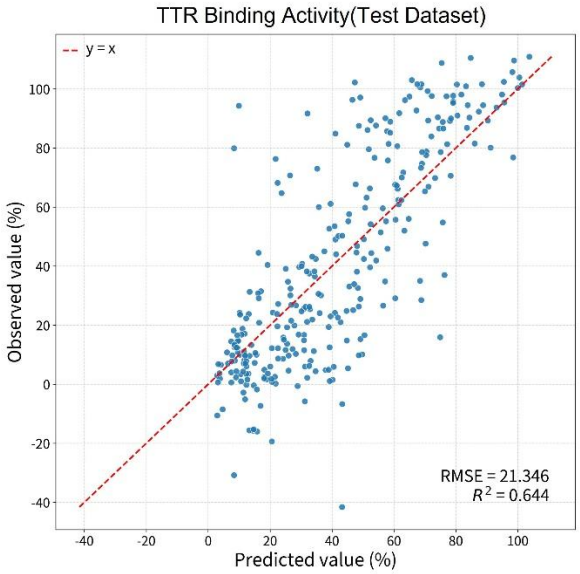

Figure 2. Scatter plot showing the relationship between predicted TTR binding activity and measured values in the test dataset using the model constructed based on input format (c). Each point represents a compound in the test dataset, and the red dashed line indicates agreement between the predicted and measured values (y = x).

 doi:10.20944/preprints202604.1199.v1

2.1.2. Setting the Applicability Domain

In this study, an applicability domain (AD) [21] was defined to determine whether the compounds to be predicted fall within the chemical space of the training dataset. Here, for each compound in the training dataset, the Euclidean distances to the five nearest neighbors (d_{i1}, ... ,d_{i5}) in the standardized (mean = 0, SD = 1) molecular descriptor space were calculated. Then, the log-transformed geometric mean was employed as the distance metric s, which is calculated as follows.

$$s_i = \log\left(\left(\prod_{j=1}^{5} d_{ij}\right)^{1/5}\right) \quad (1)$$

The maximum value of s_i in the training dataset was set as the AD threshold. Note that the s_i value was also calculated for the compounds in the test dataset. Compounds with an s_i value at or below the threshold were classified as within the AD, and those exceeding the threshold were classified as outside the AD. Normalization of the molecular descriptors for the test dataset was performed based on the distribution of the training dataset.

Here, the AD determination threshold was set to 5.15, which was the maximum s_i value in the training dataset. When the AD assessment was applied to the test dataset compounds, no compounds were classified as outside the AD (Figure 3 and Table S2).

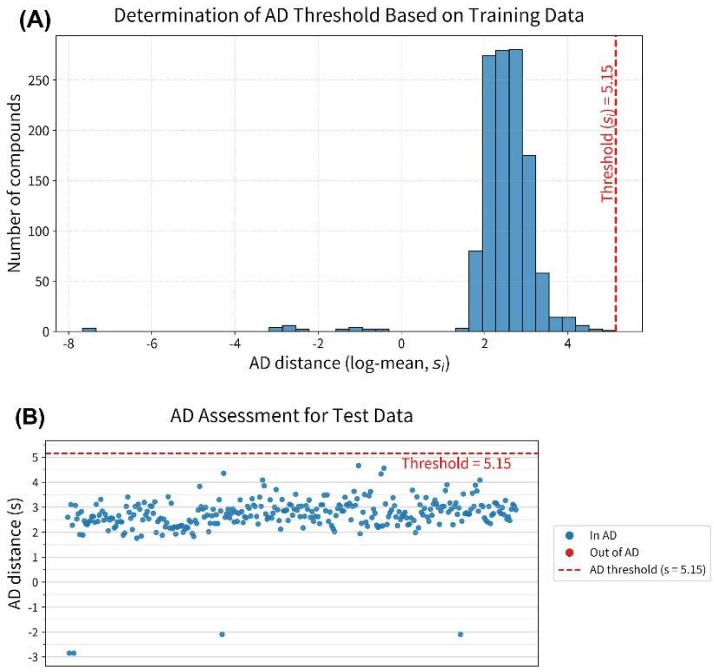

Figure 3. (**A**) Histogram of s values for each sample in the training dataset. The vertical and horizontal axes indicate the number of compounds and the s value defined as the logarithm of the geometric mean distance, respectively. (B) Relationship between the s values calculated for each compound in the test dataset and the threshold used for AD determination. All compounds were distributed below the threshold, confirming that they were within the AD.

2.1.3. SHAP Analysis

Feature importance was calculated using SHapley Additive exPlanations (SHAP) [22] to interpret the constructed model. Here, the SHAP values were computed using the TreeExplainer implemented in the Python shap library, and the entire training dataset was used for evaluation. In this study, the feature importance was calculated based on the mean absolute SHAP value for each feature and visualized using a summary plot (Figure 4). The prefixes "pre" and "post" denote the

 doi:10.20944/preprints202604.1199.v1

Mordred descriptors calculated from the structures before and after salt/complex processing, respectively. The top 10 molecular descriptors in order of importance were post_SLogP, pre_SlogP_VSA11, post_AETA_beta_ns_d, post_ETA_dEpsilon_D, post_ATSC1i, post_ETA_dEpsilon_B, pre_BertzCT, post_SMR_VSA3, post_NaasC, and post_piPC2.

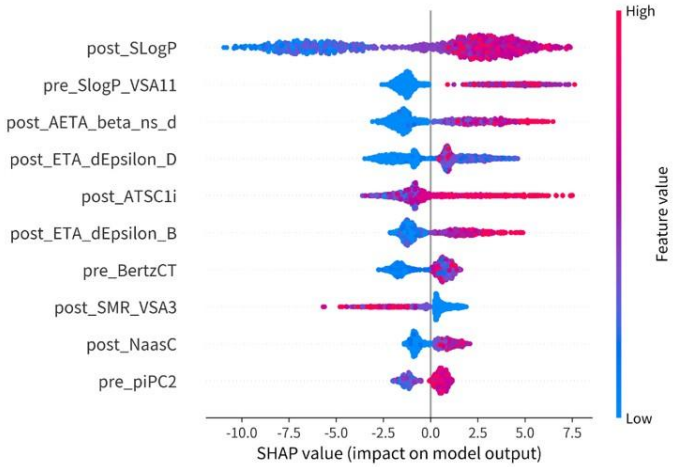

Figure 4. SHAP summary plot showing the contribution of molecular descriptors to predicted values. The horizontal axis represents the SHAP values, where positive and negative values indicate contributions toward increasing and decreasing the predicted value, respectively. Each point represents an individual compound, and the color indicates the value of the corresponding molecular descriptor, where red and blue indicate high and low values, respectively.

2.2. Correlation Analysis with Amyloid Formation Inhibition Assays

Previously reported TTR amyloid formation inhibition assays were collected from four publications, i.e., Klabunde et al. (2000), Oza et al. (2002), Adamski-Werner et al. (2004), and Palaninathan et al. (2009) [23–26]. Generally, the assay conditions were consistent across the publications (Table S3). Regarding the presentation of the results, Klabunde et al., Oza et al., and Adamski-Werner et al. reported the percentage of fibril formation (% fibril formation) calculated relative to the turbidity observed in the absence of inhibitor (set as 100%). In contrast, Palaninathan et al. reported the percentage of inhibition (% inhibition) derived from the same turbidity ratio. In the current study, to unify the metrics in a direction where a positive correlation with the model's predicted values was expected, values reported as % fibril formation were converted to inhibition rates by subtracting them from 100.

Here, 12, 13, 27, and 29 data points were obtained from each publication, respectively. Note that two compounds overlapped across the publications; thus, the experimental values from the most recently published source were retained. This deduplication process excluded three data points, resulting in a total of 78 compounds (Table S4) for analysis.

2.2.1. Correlation Analysis

For the obtained compounds, TTR binding activity predictions and AD assessment were performed using the constructed model. All compounds were classified as within the AD, and the predicted values were calculated.

The normality of the predicted TTR binding activity and amyloid formation inhibition rates was evaluated using the Shapiro–Wilk test, where W represents the test statistic. The obtained W values were 0.9438 (p = 1.824 × 10^{-3}) and 0.8487 (p = 1.933 × 10^{-7}), respectively. The null hypothesis of normality was rejected for both variables (Figure 5).

 doi:10.20944/preprints202604.1199.v1

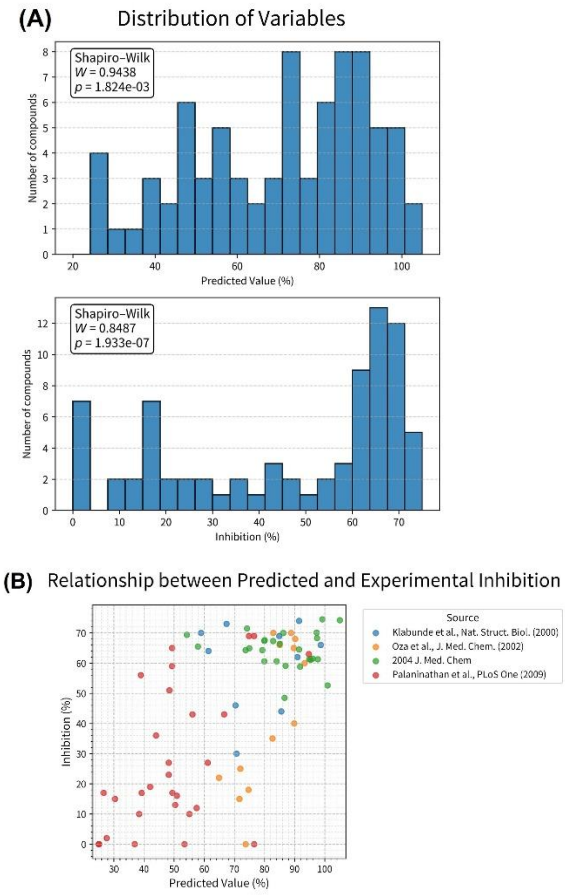

Figure 5. (**A**) Histograms showing the distributions of predicted TTR binding activity and amyloid formation inhibition rates. The upper andlower panels show the distribution of predicted TTR binding activity and amyloid formation inhibition rates, respectively. Both distributions exhibit deviations from normality. (**B**) Scatter plot showing the relationship between predicted TTR binding activity and amyloid formation inhibition rate. Each point represents a compound derived from the literature, and the colors indicate classification by literature source.

Accordingly, Spearman's rank correlation coefficient (ϱ) and the p-value were calculated to evaluate the relationship between the predicted TTR binding activity and amyloid formation inhibition rate (Figure 5). For the entire dataset, Spearman's correlation coefficient was ϱ = 0.602 (p = 5.673×10^{-9}) (Table 3). In addition, the 95% confidence interval (CI) of the Spearman's rank correlation coefficient was calculated by bootstrap resampling (1000 resamples with replacement), yielding 95% CI = 0.435–0.720. These results suggest that the observed correlation was statistically robust.

Note that variability in the correlation coefficients was observed when the Spearman's rank correlation coefficients were calculated for each publication separately (Table 3). Specifically, the study by Oza et al. (2002) exhibited a relatively high positive correlation (ϱ = 0.682, p = 1.019×10^{-2}), and that by Palaninathan et al. (2009) also exhibited a significant positive correlation (ϱ = 0.411, p = 2.670×10^{-2}). In contrast, no clear correlations were observed in the studies by Klabunde et al. (2000) and Adamski-Werner et al (2004).

 doi:10.20944/preprints202604.1199.v1

Table 3. Spearman's rank correlation coefficients for each publication and all compounds combined.

Source	N_compounds	Spearman ρ	p-value
Klabunde et al., *Nat. Struct. Biol.* (2000)	10	-0.030	9.34×10^{-1}
Oza et al., *J. Med. Chem.* (2002)	13	0.682	1.02×10^{-2}
Adamski-Werner et al., *J. Med. Chem.* (2004)	26	-0.073	7.22×10^{-1}
Palaninathan et al., *PLoS One* (2009)	29	0.411	2.67×10^{-2}
Overall	78	0.602	5.67×10^{-9}

2.3. Food Compound Screening

In this study, the PhytoHub database (version 1.4; https://phytohub.eu/, accessed on November 22, 2025), which provides Simplified Molecular Input Line Entry System (SMILES) notation for its registered compounds, was employed to screen food-derived compounds. The PhytoHub database contains structural information in SMILES notation, structural classification, information on foods from which compounds are derived, and information on compounds that are precursors or metabolites of the corresponding constituents [27].

Note that SMILES information was not assigned to all entries, and it could not be obtained for some compounds. Among the entries lacking SMILES annotations, many were conjugated compounds or compounds described by abstract names, e.g., "Blackberry flavonols," "Onion flavonols," and "Blueberry flavonols," for which the chemical structure could not be uniquely specified. Thus, to investigate the relationship between the missing SMILES information and other annotations, the status of the "Metabolites/Precursor" column was investigated across all PhytoHub entries (Table S5).

As a result, among the entries with missing "Food sources" information (n = 1750), 55.8% (977/1750) included annotations in the "Metabolites/Precursor" column. In contrast, among the entries in which both "Smiles" and "Food sources" were recorded (n = 934), the proportion was only 26.6% (248/934).

Based on these results, 383 entries with missing "Smiles" information and 1750 entries with missing "Food sources" information were excluded, and a final analytical dataset comprising 934 entries was constructed (Table S6).

The compounds in the analytical dataset were assigned to five categories, i.e., "Polyphenols," "Terpenoids," "N-containing compounds," "Miscellaneous phytochemicals," and "Not Available." Here, the "Miscellaneous phytochemicals" and "Not Available" categories, which could not be classified in an unambiguous manner, were combined as "Others."

2.3.1. Predictions for Food-derived Compounds

Predictions and AD assessment were performed for all 934 entries. Here, compounds classified as being outside the AD included five polyphenols and 24 terpenoids (Table 4). In addition, many of the compounds classified as being outside the AD were diterpenes, with 15 of the 24 outside-AD compounds falling into this category (Table S7).

The distribution of the predicted values indicated that polyphenols tended to be distributed predominantly in the high-affinity range. In contrast, terpenoids and nitrogen-containing compounds were more frequently distributed in the low-affinity range (Figure 6). The mean predicted value was highest for polyphenols at 65.4%, followed by terpenoids and nitrogen-containing compounds, at 23.6% and 24.8%, respectively.

 doi:10.20944/preprints202604.1199.v1

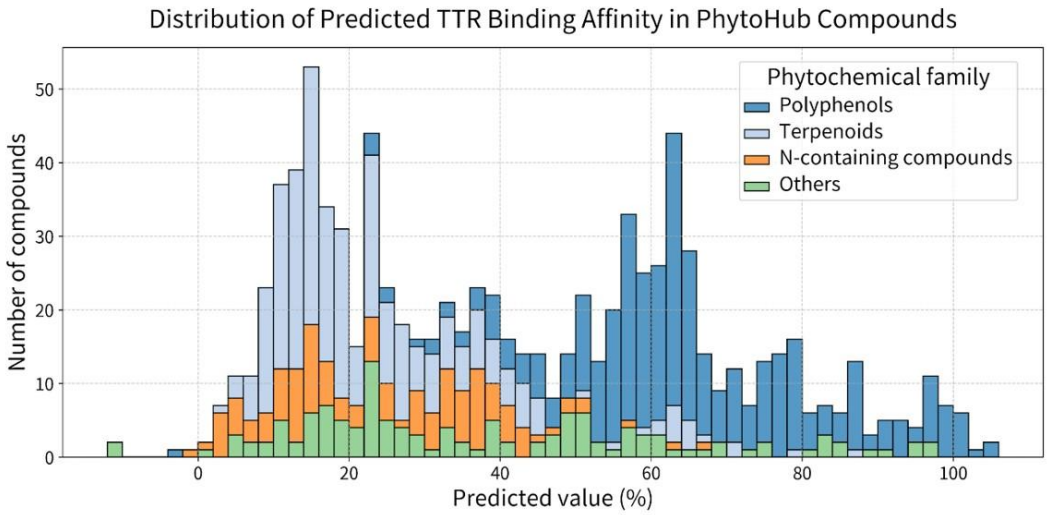

Figure 6. Histogram showing the distribution of predicted TTR binding activity values for compounds registered in PhytoHub database. The vertical and horizontal axes indicate the number of compounds and the predicted binding activity (%), respectively. The bin width was set to 2%.

Table 4. Number of entries registered in PhytoHub database by phytochemical classification and results of AD assessment.

Phytochemical family	N_compounds	AD_In	AD_Out
Polyphenols	360	355	5
Terpenoids	312	288	24
N-containing compounds	133	122	0
Others	129	129	0

2.3.2. Extraction of Candidate Compounds

To identify compounds with high binding activity among those classified as within the AD, a threshold of 85% for the predicted value was set, and the compounds meeting or exceeding this threshold were extracted as candidate compounds. Note that this threshold value was based on criteria utilized in previous study for candidate compound selection; however, it does not represent a definitive biological boundary [14].

To investigate the effect of threshold selection, the number of candidate compounds was compared across thresholds of 75%–100%. The number of candidate compounds decreased monotonically with an increasing threshold value. At the 85% threshold, 63 compounds (6.96% of the 905 compounds within the AD) were extracted (Table 5).

Table 5. Effect of threshold selection on the number of candidate compounds.

Threshold	N_compounds	Proportion (%)
70%	138	15.25%
75%	111	12.27%
80%	76	8.40%
85%	63	6.96%
90%	41	4.53%
95%	30	3.31%
100%	9	0.99%

This threshold was considered appropriate because it avoided an excessively large number of candidates while maintaining a sufficient pool for exploration. Thus, the 85% threshold, which

 doi:10.20944/preprints202604.1199.v1

balanced the number of candidates and selectivity in line with previous criteria, was adopted as the basis for candidate compound extraction.

Based on this criterion, 63 candidate compounds were extracted (Table 6), of which 54 were classified as polyphenols (Table S7).

Table 6. Candidate compounds extracted using 85% predicted value threshold.

RANK	Chemical Name	Phytochemical family	Predicted value
1	Glycitein	Polyphenols	104.67
2	Daidzein	Polyphenols	104.02
3	2'-Hydroxydaidzein	Polyphenols	102.79
4	Cyanidin	Polyphenols	101.71
5	Genistein	Polyphenols	101.69
6	Hispidulin	Polyphenols	101.55
7	Apigenin	Polyphenols	101.42
8	Calycosin	Polyphenols	101.32
9	Biochanin A	Polyphenols	100.11
10	Herbacetin	Polyphenols	99.89
11	Delphinidin	Polyphenols	99.76
12	Pratensein	Polyphenols	99.68
13	Quercetin	Polyphenols	99.31
14	Resveratrol (cis-)	Polyphenols	98.69
14	Resveratrol (trans-)	Polyphenols	98.69
16	Luteolin	Polyphenols	98.05
17	Betuletol	Polyphenols	97.79
18	Isorhamnetin	Polyphenols	97.77
19	Rhapontigenin	Polyphenols	97.42
20	Prunetin	Polyphenols	97.40
21	Isorhapontigenin	Polyphenols	97.30
22	Petunidin	Polyphenols	97.26
23	Pinosylvin	Polyphenols	97.06
24	Kaempferol	Polyphenols	96.88
25	Pelargonidin	Polyphenols	96.76
26	1,3-cis-Tetrahydroxyphenylindan	Others	96.59
26	1,3-trans-Tetrahydroxyphenylindan	Others	96.59
28	Piceatannol (cis-)	Polyphenols	95.17
28	Piceatannol	Polyphenols	95.17
30	desmethylxanthohumol	Others	95.02
31	Emodin	Others	94.92
32	Oxyresveratrol (cis-)	Polyphenols	93.61
33	Oxyresveratrol	Polyphenols	93.56
34	Dihydro-resveratrol	Polyphenols	93.14
35	Formononetin	Polyphenols	92.92
36	Liquiritigenin	Polyphenols	92.18
37	Dihydro-piceatannol	Polyphenols	91.59
38	Chrysophanol	Others	91.06
39	Naringenin	Polyphenols	90.90
40	Myricetin	Polyphenols	90.66
41	Hesperetin	Polyphenols	90.25
42	Iriflophenone	Polyphenols	89.39
43	Ellagic acid	Polyphenols	88.86
44	Rhein	Others	88.08
45	Xanthohumol	Polyphenols	87.80
46	Thunberginol C	Polyphenols	87.75
47	Maclurin	Polyphenols	87.64
48	Phloretin	Polyphenols	87.33
49	Scirpusin A	Polyphenols	87.25
50	Tectochrysin	Polyphenols	87.17
51	Gnetin C	Polyphenols	87.09
52	5,5'-Diferulic acid	Polyphenols	87.00
53	Eugenol	Polyphenols	86.76
54	Viniferin (cis-Epsilon-)	Polyphenols	86.65
54	Viniferin (Epsilon-)	Polyphenols	86.65
54	Viniferin (Omega-)	Polyphenols	86.65
57	Farnesol	Terpenoids	86.04
58	Falcarindiol	Others	85.95
59	Pseudobaptigenin	Polyphenols	85.65
60	6-Prenylnaringenin	Polyphenols	85.35
61	Parthenocissin A	Polyphenols	85.30
61	Quadrangularin A	Polyphenols	85.30
63	Tragopogonic acid	Others	85.23

Note that the PhytoHub database is partially constructed based on manual registration by domain experts; thus, some compounds may not be clearly assignable to standard scientific classifications [27].

2.3.3. Extraction of Foods Containing Candidate Compounds

Tabulation was stratified by the number of candidate compound types contained in each food. Here, to assess the trend in compound recording across foods, the total number of compounds listed in the PhytoHub database, including those classified as outside the AD, was also tabulated.

As a result, candidate compounds were found in 126 food sources (Table 7). The food source containing the largest number of candidate compound types was red wine, which contained 14 compounds with high predicted binding activity. In addition to red wine, other wines (white and rosé) and beverages, e.g., coffee and fruit juices, including apple and peach juices,, were also ranked highly. When limited to plant foods, the highest number of candidate compound types was found in red clover, followed by garden rhubarb. These results may have been influenced by the number of compounds recorded for each food and the recording tendencies of the database.

 doi:10.20944/preprints202604.1199.v1

Table 7. Food sources containing candidate compounds, organized by the number of candidate compound types and database compound information for each food. N_compound indicates the number of candidate compounds contained in each food source. Values in parentheses indicate the number of compounds classified as outside the AD among all compounds listed for each food in PhytoHub.

N_compound	Food sources
14	Red wine(29)
9	White wine(19)
8	Red clover(10)
7	Rosé wine(9)
6	Garden rhubarb(7) Black tea(39)
5	rhubarb(6) Chickpea(10) Apple juice(25) Beer(39) Apple(57)
4	Acerola(4) Blueberry(21) Green tea(26) Grape wine(43) Tomato(44) Strawberry(52) Broccoli(58) Coffee(59)
3	Camomile tea(3) Ceylon tea(3) Dessert wine(3) Peach juice(3) Red tea(3) Tilia tea(3) Red wine grape(4) White wine grape(4) Grapes(7) Cranberry(8) European cranberry(9) Soy bean(19) Common oregano(22) Blackberry(27) American cranberry(29) Onion(33)
2	Mulberry(2) Soy milk(2) Angelica(3) Common salsify(3) Deerberry(3) Pomace(3) Red champagne(3) Red grape juice(4) Pomegranate(5) Tart cherry(5) Bilberry(6) Cranberry juice(6) Liquorice(6) rhubarb root(7) Celery stalks(9) Parsnip(10) Common bean(11) Cherry tomato(12) Pistachio(12) Blackcurrant(13) Common sage(13) Common thyme(13) Parsley(13) Peanut(13) Red raspberry(13) Mango(14) Almond(17) Globe artichoke(19) Common cabbage(21) Orange juice(25) Sweet orange(25) Olive, black(34)
1	Cagaita(1) European chestnut(1) Grumixama juice(1) Jabuticaba(1) Jackfruit(1) Mexican origano(1) Partridge berry(1) Radicchio(1) Sparkleberry(1) Cassia cinnamon(2) Chervil(2) Cloves(2) Cocoa liquor(2) Itadori tea(2) White champagne(2) Grape juice(3) Chamomile tea(4) Coriander(4) Lingonberry(4) Caraway(5) Celery leaves(5) Chicory(5) Grapefruit juice(6) Nutmeg(6) Sweet marjoram(6) Broad bean(7) Fennel(7) Lime(7) Cinnamon(8) Pecan nut(8) Sweet basil(8) Whisky(8) Lemon(9) Lentils(9) Pomegranate juice(10) Apple cider(11) Barley(11) Kiwi(11) Oat(11) Chocolate(12) Common pea(12) Corn(12) Dill(12) Common wheat(13) Common walnut(14) Celeriac(15) Peppermint(15) Grapefruit(18) Flaxseed(19) Rosemary(20) Peach(21) Banana(26) Carrot(28) Lettuce(30) Apricot(31) Olive, green(32) Grape(63)

3. Discussion

3.1. Prediction Model Construction and Accuracy

In the comparison of the three feature input formats examined during the model construction process, format (c), which integrated molecular descriptors calculated from both the original and salt/complex-processed structures, exhibited the highest predictive accuracy, as shown in Table 2.

This suggests that integrating information derived from different structural states before and after processing may be useful for the target prediction task.

However, it could not be clearly determined whether the improvement in performance was attributable to the contribution of information derived from the salt/complex-processed structures or the overall effect of expanding the feature representation through integration. Thus, the underlying mechanism remains uncertain.

In addition, the RMSE of format (c) was comparable to that of the top-ranking models (corresponding to seventh place) in the published Tox24 Challenge results (Figure 2) [28]. Furthermore, considering that teams whose prediction scores were not significantly different from the winning model according to a paired t-test were designated as “group winners,” and that the top 11 teams fell into this category, the model constructed in the current study is considered to have similar predictive performance [20].

Note that no compounds in the test dataset were classified as outside the AD in this study (Figure 3), which suggests that the training dataset covered a relatively broad chemical space and that the test dataset was contained within this range.

3.2. Interpretation by SHAP

Among the top 10 descriptors identified by the SHAP analysis, both pre- and post-salt/complex processing descriptors were included (Figure 4). The top two descriptors were related to lipophilicity, and higher values contributed positively (toward higher binding activity) to the predicted values (Table 8). This is consistent with the hydrophobic environment of the TTR binding site [29]. In addition, several descriptors related to aromaticity and unsaturation were observed (Table 8). Considering that T4, the endogenous ligand of TTR, contains aromatic rings, the model appears to capture hydrophobic interactions arising from aromatic rings as features contributing to binding

 doi:10.20944/preprints202604.1199.v1

activity. Furthermore, the ETA_dEpsilon_D descriptor reflects the presence and number of hydrogen atoms bonded to heteroatoms, which suggests that the model may incorporate the hydrogen bond donating capacity of hydroxyl groups into its predictions (Table 8). Crystal structure analyses have reported that the hydroxyl groups of flavonoids can participate in hydrogen bond formation, including water-mediated interactions, and the results of the current study are consistent with this observation [30]. Taken collectively, the results indicate that the model learned and reflected structural features that are relevant to TTR binding activity in a manner that is consistent with existing knowledge.

Table 8. Descriptions of top 10 Mordred molecular descriptors identified as most important by SHAP analysis.

Descriptor	Mordred official description
SLogP	Wildman-Crippen LogP
SlogP_VSA11	MOE logP VSA Descriptor 11 (0.50 ≤ x < 0.60)
AETA_beta_ns_d	averaged nonsigma contribution to valence electron mobile count
ETA_dEpsilon_D	ETA delta epsilon (type: D)
ATSC1i	centered Moreau–Broto autocorrelation of lag 1 weighted by ionization potential
ETA_dEpsilon_B	ETA delta epsilon (type: B)
BertzCT	Bertz CT
SMR_VSA3	MOE MR VSA Descriptor 3 (1.82 ≤ x < 2.24)
NaasC	number of aasC
piPC2	2-ordered pi-path count (log scale)

3.3. Correlation Between Amyloid Formation Inhibition and Predicted TTR Binding Activity

Note that searching for TTR stabilizer candidates differs from the original purpose of the assay data provided in the Tox24 Challenge [17]. Thus, it was necessary to verify whether the predicted values were related to amyloid formation inhibition.

Here, the compounds identified in the TTR amyloid formation inhibition assays reported by Klabunde et al. (2000), Oza et al. (2002), Adamski-Werner et al. (2004), and Palaninathan et al. (2009) were integrated, and the normality of the predicted TTR binding activity and amyloid formation inhibition rates was evaluated using the Shapiro–Wilk test. The results demonstrated that neither variable exhibited normality (Figure 5).

The lack of normality in the amyloid formation inhibition rates may be attributable to structural bias in the compound sets used in each publication. Although data from multiple publications were integrated, many compounds were derivatives generated by systematic structural modification of lead compounds with expected amyloid formation inhibitory activity, which may have introduced bias in the compound properties [23–26]. Similarly, the predicted values may have been influenced by the structural bias of the assayed compounds.

Spearman's rank correlation coefficient was calculated for the correlation analysis. Here, the calculated result was $\rho = 0.602$ ($p = 5.673 \times 10^{-9}$), which indicates a positive correlation between the predicted TTR binding activity and amyloid formation inhibition rate (Table 3). In addition, the 95% CI of the Spearman's rank correlation coefficient calculated by bootstrap resampling (1000 resamples with replacement) was 0.435–0.720. In contrast, variability in the correlation coefficients was observed across publications (Table 3). Specifically, positive correlations were observed in the publications by Oza et al. (2002) and Palaninathan et al. (2009), whereas no clear correlations were observed in the publications by Klabunde et al. (2000) and Adamski-Werner et al. (2004). This suggests that the correlation observed in the current study was not uniform across all publications and may have been driven by trends in certain publications. However, each publication included derivatives based on the same lead compounds; thus, the present results were not based on completely independent compound sets.

 doi:10.20944/preprints202604.1199.v1

In this study, structural diversity was ensured by integrating compounds reported across multiple publications, and the relationship between model predicted values and inhibitory activity was evaluated. The results indicate that while the predicted TTR binding activity does not fully explain amyloid formation inhibitory activity, it serves as a reasonable index with a certain degree of association with inhibitory activity. Thus, screening using this model is considered useful as an exploratory approach to narrow down candidate compounds with amyloid formation inhibitory activity from structurally diverse compound sets.

3.4. PhytoHub Database

As mentioned previously, the PhytoHub database contains missing SMILES information (383 entries) and missing food source information (1750 entries) out of 2696 total entries (Table S5). The high prevalence of conjugated compounds among entries without SMILES information may be because conjugation can occur at multiple sites, thereby making it difficult to uniquely define the chemical structure. In addition, compounds described by abstract names cannot be uniquely mapped to specific chemical structures, which likely explains the absence of the SMILES information.

Furthermore, the high proportion of precursor/metabolite annotations among entries with missing food source information suggests that these compounds may be organized as metabolic products generated in vivo rather than compounds directly present in food.

The analysis performed in this study was limited to compounds with clear structural and food source information. Thus, the resulting dataset may be biased toward parent compounds directly present in food, and this point should be noted as a constraint regarding the AD and interpretation of this screening.

In addition, while compounds with TTR stabilizing activity may lose their activity through conjugation metabolism, some conjugated metabolites, e.g., the sulfate conjugate of resveratrol, have been reported to retain their activity [31]. Thus, future studies must collect the structures of conjugates and metabolites in addition to parent compounds and apply the same prediction model employed in the current study to perform evaluations that more closely reflect actual in vivo activity.

3.5. Compounds with High Predicted Values

Setting the predicted value threshold to 85% resulted in the extraction of 63 compounds (Table 6). Among these compounds, flavonoid compounds were particularly abundant (Table S7). This trend is consistent with previous studies reporting that polyphenols exhibit affinity for the TTR binding pocket and can be interpreted as suggesting that the present model reflects, to some extent, molecular structural features associated with TTR binding [30–35].

3.6. Foods Containing Candidate Compounds

In this study, red wine was found to contain the largest number of candidate compounds. In addition, other beverages, including different types of wines (white and rosé as well as red) and juices were among the foods with the highest number of candidate compound types (Table 7). However, even within the same category, products may differ in terms of their raw materials and manufacturing conditions, and products with different characteristics may have been aggregated under a single entry. As a result, the number of candidate compound types may have been overestimated. Thus, the superiority of individual foods cannot be directly evaluated based solely on the number of candidate compound types, and all foods identified in the present analysis should be regarded only as candidates. In addition, this analysis did not consider the content levels of each component; thus, additional investigations based on content levels are required to evaluate the actual contribution of ingestion. In this context, the number of candidate compound types may serve as a useful reference to prioritize future investigations. Notably, the foods presented in the current study are merely candidates based on the compound information in the database. In other words, they do not constitute recommendations for the consumption of specific foods. In particular, alcohol

 doi:10.20944/preprints202604.1199.v1

consumption is a major risk factor for disease burden, and the associated health risks increase dose-dependently [36]. Therefore, even for foods containing alcohol, the health risks associated with exposure must be considered.

3.7. Study Limitations

3.7.1. Quality of the Training Dataset

To improve predictive accuracy, it is important to ensure data quality. The assay addressed in this study targets the TTR binding pocket; thus, it may be influenced by activity cliffs, in which subtle structural changes lead to drastic changes in activity [37]. For example, even a slight modification, e.g., replacing an ether oxygen atom with a secondary amine, may affect activity considerably.

To adequately capture such activity cliffs, it is necessary to collect data covering a wide range of structurally similar compounds and estimate trends statistically. Thus, the sample size of the data used in the current study may have been insufficient to fully reflect compound diversity.

In addition, the dataset may contain compounds with low measurement reliability. When analogous compounds differing in terms of only the number of carbon atoms in the main chain were extracted from the training data for comparison with compounds in the test dataset, irregular fluctuations were observed in the measured values, even though a monotonic change with carbon number would be expected (Figure S1). This suggests that the dataset may contain anomalous values originating from measurement failures or related issues. If such inconsistent data are mixed into the training dataset, the prediction accuracy obtained on the test dataset may be reduced.

Note that the measured values in the present data represent the averages of triplicate measurements, and variability between measurements is expected [14]. Considering this variability, constructing models using data with improved measurement reliability, which can be achieved by increasing the number of replicates to reduce the influence of random errors, could potentially yield improved predictive accuracy.

3.7.2. Regarding Screening

In this study, the screening process was based on the predicted TTR binding activity values and did not evaluate actual binding activity directly. Although predictions within the AD are expected to have a certain level of reliability, they do not guarantee quantitative agreement between the predicted and measured values. Thus, further experimental validations of the results are required.

In addition, the 85% predicted value threshold employed for the candidate compound extraction process lacks a definitive biological basis. As a result, compounds with actual activity may have been excluded from the analysis.

Furthermore, this study did not consider pharmacokinetic factors, e.g., the content of compounds in foods, bioavailability, and blood kinetics. In particular, polyphenols are known to undergo degradation by gut microbiota and conjugation metabolism in vivo, and they are present in the blood mainly as metabolites [38]. Thus, the predicted values for the parent compounds evaluated in this study do not necessarily directly reflect actual TTR binding capacity or stabilization activity in vivo.

Accordingly, the results obtained in this study are presented as an exploratory screening to identify compounds with potential TTR stabilization activity. Future studies should include experimental validation of candidate compounds, as well as extensive evaluations that incorporate metabolites and consider pharmacokinetic properties.

4. Materials and Methods

The model construction and statistical analyses in this study were performed in a Python (version 3.8.0) environment.

 doi:10.20944/preprints202604.1199.v1

4.1. Dataset for Model Construction

The train, leaderboard, and blind data used for model construction were obtained from OCHEM (https://ochem.eu/static/challenge-data.do, accessed on May 29, 2025). These data were provided as assay data in the 2024 Tox24 Challenge.

The Tox24 Challenge is a competition aimed at evaluating advances in computational methods to predict the in vitro activity of chemical compounds. Each data entry includes the compound's SMILES structure and TTR binding activity assay value [17].

In this study, the train (N = 1012) and leaderboard (N = 200) datasets were used for model training, and the blind (N = 300) dataset was used to evaluate the performance of the constructed model. The measured values were obtained using the TTR-ANSA fluorescence assay, which utilizes the change in fluorescence intensity when the fluorescent probe ANSA bound to TTR is displaced by a compound. The compound activity was calculated based on a standard curve of T4, with high-concentration T4 set to 100% activity (ANSA completely displaced from TTR) and low-concentration T4 set to 0% activity (ANSA not displaced). In addition, each compound was measured at a single concentration, and the TTR binding activity was calculated from the resulting fluorescence intensity [14].

The data provided by OCHEM were used without modification in this study. In other words, no compounds were added or removed, and no corrections or outlier removal was applied to the measured values.

4.2. Salt/Complex Processing and Molecular Descriptor Calculation

For SMILES notation recorded as salts or complexes, processing was performed to extract only the active moiety. SMILES data were split into fragments using periods as delimiters, and the main component was selected from each set of fragments. In this study, the fragment with the largest number of atoms was extracted as the main component. Then, charge neutralization was applied to the extracted structure, and canonical SMILES notation were regenerated using RDKit (version 2024.03.2). Details about the processing procedure are shown in Figure S2.

Furthermore, molecular descriptors were calculated using Mordred (version 1.2.0). For simplicity, only 2D descriptors were calculated without generating 3D structures from SMILES (ignore_3D=True). Then, descriptors containing missing values, those with zero variance, and duplicate descriptors were excluded. Three input formats were defined for molecular descriptors, i.e., (a) using only descriptors calculated from structures without salt/complex processing, (b) using only descriptors from structures after salt/complex processing, and (c) integrating descriptors from both preprocessing and postprocessing structures. In format (b), the salt/complex processing caused structural duplication among some compounds; thus, duplicate compounds were merged. In this case, the activity value was represented by the median. As a result, the number of training compounds decreased from 1212 to 1186. Detailed lists of the descriptors and excluded items corresponding to each input format are given in Tables S8–S10.

4.3. Model Construction

The model was constructed using LightGBM (version 3.3.5) in a two-stage process comprising feature selection and hyperparameter optimization. Here, feature selection was performed based on five-fold CV, where the data were divided into five subsets, with one subset used for validation and the remaining four subsets used for training in each fold. In each fold, a LightGBM model was trained using only the training dataset, and feature importance based on the number of splits was calculated. Hyperparameters were fixed during this stage. The obtained importance scores were averaged across folds, and the features with a mean importance greater than 0 were retained. All subsequent training and evaluation processes used the selected feature set. The feature selection results for each input format are shown in Tables S8–S10.

 doi:10.20944/preprints202604.1199.v1

Hyperparameter optimization was performed in stages. First, Bayesian optimization using Optuna was applied to identify the optimal combination of learning_rate, num_leaves (tree complexity), min_data_in_leaf, min_sum_hessian_in_leaf (leaf constraints), bagging_fraction, feature_fraction (subsampling parameters), lambda_l1, and lambda_l2 (regularization terms). Early stopping was applied during this stage, and training was terminated when the validation metric ceased to improve. Next, seven candidate values were set for the number of iterations (n_estimators) by varying the optimal iteration count obtained in the previous stage by several hundred in both directions, and the final value was determined by grid search. Finally, the mean prediction accuracy across five-fold CV was compared for the input formats (a), (b), and (c), and the format that obtained the highest performance was selected. The final model was then constructed using the entire training dataset based on the selected format. The search ranges and determined hyperparameters are shown in Table S11.

4.4. Final Model Evaluation

As the evaluation metric, the RMSE, representing the square root of the mean of squared differences between the predicted and measured values for each sample, was calculated. Note that the RMSE is scale-dependent on the objective variable; thus, the coefficient of determination (R^2) was also calculated as a supplementary metric.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{2}$$

In addition, to evaluate the reliability of the predictions, an AD was defined to determine whether each compound fell within the chemical space of the training dataset. Here, the molecular descriptors were normalized based on the training dataset, and the same transformation was applied to the test dataset. For each sample, the Euclidean distances to the five nearest neighbors in the training dataset were calculated, and the AD index was derived from the log-transformed geometric mean of these distances.

4.5. Model Interpretation by SHAP

The SHAP method is used to quantitatively evaluate the contribution of each feature to model predictions. The explanatory model is expressed as follows:

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M}\phi_i x_i' \tag{3}$$

where ϕ_0, ϕ_i, and x_i' denote the baseline value, the contribution of each feature, and the presence or absence of feature i in the simplified input, respectively. The contribution of each feature is defined as a Shapley value [22].

In this study, the SHAP values were calculated using the training dataset for model interpretation. For missing features, the influence was evaluated by marginalization using the training dataset as the background distribution. Then, the feature importance was evaluated based on the obtained SHAP values, and the top 10 features with the largest contributions were visualized.

4.6. Literature on TTR Amyloid Formation Inhibition Assays

Compounds and assay data were collected from four publications reporting TTR amyloid formation inhibition assays, i.e., Klabunde et al. (2000), Oza et al. (2002), Adamski-Werner et al. (2004), and Palaninathan et al. (2009). In each publication, the assays were based on inducing TTR amyloid formation under acidic conditions and evaluating fibril formation (or inhibition rate) by turbidity measurement.

Klabunde et al. (2000) evaluated the amyloid formation rate of wild-type TTR at concentrations of 7.2, 3.6, and 1.8 μM for 12 compounds classified primarily as nonsteroidal anti-inflammatory drugs

 doi:10.20944/preprints202604.1199.v1

[23]. Oza et al. (2002) evaluated 13 diclofenac analogs at 3.6 μM using wild-type and mutant TTR (V30M, L55P) [24], and Adamski-Werner et al. (2004) evaluated 20 compounds, primarily diflunisal analogs, at 7.2 and 3.6 μM for amyloid formation rate against wild-type TTR [25]. Finally, Palaninathan et al. (2009) evaluated 29 compounds, primarily β-aminoxypropionic acid derivatives, at 3.6 μM for amyloid formation inhibition rate against wild-type TTR [26].

Although the assay conditions differed in part across these publications, data at 3.6 μM against wild-type TTR were available in all publications. Thus, the dataset used in the current study was standardized under this condition.

Furthermore, SMILES notations were not provided in the publications; thus, the chemical structures were drawn using ChemDraw (Version 25.0.2.14) to obtain the SMILES notations for each compound.

In each publication, the amyloid formation and inhibition rates were defined relative to the amyloid formation in the control without an inhibitor. To unify these metrics, the amyloid formation rate was converted to an inhibition rate by subtracting it from 100%, and the dataset was constructed accordingly (Table S4).

Then, the predictions and AD assessment were performed on the acquired dataset. Compounds classified as outside the AD were excluded from the subsequent correlation analysis.

4.7. Correlation Analysis

To evaluate the association between the predicted TTR binding activity and amyloid formation inhibition rate, the normality of each variable was evaluated using the Shapiro–Wilk test. If normality was confirmed, Pearson's correlation coefficient was considered. However, if normality was not established, Spearman's rank correlation coefficient, which is a nonparametric method, was employed to evaluate the association between the two variables. Here, the correlation coefficient (ϱ) and p-value were calculated.

In addition, to evaluate the uncertainty of Spearman's rank correlation coefficient, the 95% CI was calculated via bootstrap resampling (1000 resamples with replacement). To assess variability across the publications, Spearman's rank correlation coefficient was calculated separately for each publication.

4.8. PhytoHub Database

PhytoHub is an online database that comprehensively catalogs phytochemicals present in foods commonly consumed by humans. In addition to information dispersed across more than 20 existing databases, the PhytoHub database integrates data manually extracted from the literature by experts and data obtained experimentally on collaborative research platforms.

PhytoHub was designed for research on nutritional metabolomics and the health effects of food-derived compounds, and it provides information on food-derived phytochemicals and their metabolites. Searches can be performed based on foods, compound names, molecular formulas, and related information [27].

PhytoHub version 1.4 contains approximately 1200 phytochemicals (including polyphenols, terpenoids, and alkaloids) present in more than 350 foods, together with more than 560 related human or animal metabolites.

PhytoHub (version 1.4; https://phytohub.eu/, accessed on November 22, 2025) was obtained. In this study, the "Chemical Name," "Phytochemical family," "Smiles," "Food sources," and "Metabolites / Precursor" columns were extracted from the PhytoHub database for analysis (Table S5).

First, because SMILES information was not assigned to all entries, the status of the "Food sources" and "Metabolites / Precursor" columns was investigated across all PhytoHub entries to evaluate the relationship between the missing SMILES information and other annotations. Specifically, the annotation rates of "Metabolites / Precursor" were calculated and compared between entries with missing "Food sources" information and entries with "Smiles" and "Food sources"

 doi:10.20944/preprints202604.1199.v1

information. Then, entries with missing "Smiles" or "Food sources" information were excluded to construct the analytical dataset (Table S6).

The compounds in the analytical dataset were classified based on the "Phytochemical family" column. However, entries classified as "Miscellaneous phytochemicals" and "Not Available" were combined as "Others" because their classification could not be uniquely determined.

Predictions and AD assessment were then performed, and the compounds within the AD with predicted values exceeding 85% were extracted as candidate compounds. Food sources corresponding to these candidate compounds were then extracted based on the "Food sources" information.

5. Conclusions

ATTRwt is a progressive disease in which amyloid deposition in the heart is frequently observed in elderly individuals [16]. However, unlike lifestyle-related diseases, preventive intervention strategies other than pharmaceutical agents have not been sufficiently established. Thus, this study investigated food candidates for potential preventive intervention in ATTRwt using in silico methods.

In this study, a TTR binding activity prediction model was constructed and applied to screen registered compounds in PhytoHub, a phytochemical database. As a result, food candidates containing compounds with high predicted TTR binding activity were identified.

The significance of this study lies in presenting compounds and food candidates with potential relevance to ATTRwt prevention, thereby expanding the scope for future experimental validation and clinical investigation. However, the model constructed in this study does not directly predict TTR stabilization activity but rather evaluates TTR binding activity. Thus, whether these compounds can in fact stabilize TTR requires extensive verifications at the compound level and clinical evaluations.

The results of this study are positioned as exploratory findings for considering intervention candidates for ATTRwt. Thus, multifaceted investigations, including experimental validations and clinical evaluations, are required in the future.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org. Figure S1. Example of a compound with suspected measurement failure. For nonanal compounds included in the test dataset, analogous compounds differing only in the number of main-chain carbon atoms were extracted from the training dataset and compared. The measured value of octanal did not fall within the range bounded by the measured values of heptanal and nonanal. Figure S2. Pseudocode for salt/complex processing. Table S1. Model construction dataset containing SMILES, TTR binding activity (%), measurement units, and dataset assignment (train/leaderboard datasets). Table S2. Test dataset with predicted TTR binding activity values, AD scores, AD thresholds, and AD classification results. Table S3. Experimental conditions for the TTR amyloid aggregation inhibition assays used for dataset construction. Table S4. Amyloid formation inhibition assay dataset compiled from four literature sources, including SMILES, % inhibition, predicted TTR binding activity values, and AD assessment results. Table S5. Complete PhytoHub dataset extracted for analysis, including "Chemical Name," "Phytochemical family," "Smiles," "Food sources," and "Metabolites / Precursor." Table S6. PhytoHub analytical dataset after exclusion of entries with missing "Smiles" or "Food sources" information. Table S7. PhytoHub analytical dataset with predicted TTR binding activity values, AD scores, AD thresholds, and AD classification results.

Table S8. Detailed descriptor list and preprocessing/feature selection results for input format (a) using descriptors calculated from structures without salt/complex processing. Table S9. Detailed descriptor list and preprocessing/feature selection results for input format (b) using descriptors calculated from structures after salt/complex processing. Table S10. Detailed descriptor list and preprocessing/feature selection results for input format (c), integrating descriptors from both preprocessing and postprocessing structures. Table S11.

 doi:10.20944/preprints202604.1199.v1

Search ranges and selected hyperparameters for the LightGBM models constructed based on each input format.

Author Contributions: Conceptualization, Y.U.; methodology, Y.U.; software, Y.U.; validation, Y.U., Y.I.; formal analysis, Y.I. and Y.U.; investigation, Y.I. and Y.U.; resources, Y.I. and Y.U.; data curation, Y.I. and Y.U.; writing—original draft preparation, Y.I..; writing—review and editing, Y.I. and Y.U.; visualization, Y.I. and Y.U.; supervision, Y.U.; project administration, Y.U.; funding acquisition, Y.U. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: Data are contained within the article and Supplementary Materials.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AD applicability domain
CI confidence interval
CV cross-validation
RMSE root mean square error
SD standard deviation
SMILES Simplified Molecular Input Line Entry System

References

1. Ando, Y.; Adams, D.; Benson, M.D.; Berk, J.L.; Planté-Bordeneuve, V.; Coelho, T.; Conceição, I.; Ericzon, B.-G.; Obici, L.; Rapezzi, C.; et al. Guidelines and New Directions in the Therapy and Monitoring of ATTRv Amyloidosis. *Amyloid* **2022**, *29*, 143–155, doi:10.1080/13506129.2022.2052838.
2. Saraiva, M.J.M.; Costa, P.P.; Goodman, D.S. Biochemical Marker in Familial Amyloidotic Polyneuropathy, Portuguese Type. Family Studies on the Transthyretin (Prealbumin)-Methionine-30 Variant. *J Clin Invest* **1985**, *76*, 2171–2177, doi:10.1172/JCI112224.
3. Jacobson, D.R.; Pastore, R.D.; Yaghoubian, R.; Kane, I.; Gallo, G.; Buck, F.S.; Buxbaum, J.N. Variant-Sequence Transthyretin (Isoleucine 122) in Late-Onset Cardiac Amyloidosis in Black Americans. *New England Journal of Medicine* **1997**, *336*, 466–473, doi:10.1056/NEJM199702133360703.
4. Westermark, P.; Sletten, K.; Johansson, B.; Cornwell, G.G. Ⅲ Fibril in Senile Systemic Amyloidosis Is Derived from Normal Transthyretin. *Proceedings of the National Academy of Sciences* **1990**, *87*, 2843–2845, doi:10.1073/pnas.87.7.2843.
5. Hammarström, P.; Jiang, X.; Hurshman, A.R.; Powers, E.T.; Kelly, J.W. Sequence-Dependent Denaturation Energetics: A Major Determinant in Amyloid Disease Diversity. *Proceedings of the National Academy of Sciences* **2002**, *99*, 16427–16432, doi:10.1073/pnas.202495199.
6. Lai, Z.; Colón, W.; Kelly, J.W. The Acid-Mediated Denaturation Pathway of Transthyretin Yields a Conformational Intermediate That Can Self-Assemble into Amyloid. *Biochemistry* **1996**, *35*, 6470–6482, doi:10.1021/bi952501g.
7. Colon, W.; Kelly, J.W. Partial Denaturation of Transthyretin Is Sufficient for Amyloid Fibril Formation in Vitro. *Biochemistry* **1992**, *31*, 8654–8660, doi:10.1021/bi00151a036.
8. Anan, I. Advances in the Treatment of Transthyretin Amyloidosis. *egastro* **2025**, *3*, doi:10.1136/egastro-2025-100198.
9. Kazi, D.S.; Bellows, B.K.; Baron, S.J.; Shen, C.; Cohen, D.J.; Spertus, J.A.; Yeh, R.W.; Arnold, S.V.; Sperry, B.W.; Maurer, M.S.; et al. Cost-Effectiveness of Tafamidis Therapy for Transthyretin Amyloid Cardiomyopathy. *Circulation* **2020**, *141*, 1214–1224, doi:10.1161/CIRCULATIONAHA.119.045093.

 doi:10.20944/preprints202604.1199.v1

10. Fukunari, A.; Matsushita, H.; Furukawa, T.; Matsuzaki, H.; Tanaka, H.; Ogawa, Y.; Sugimura, Y.; Inoue, F.; Ueda, M.; Ando, Y. Arginine: A Potential Prophylactic Supplement for Transthyretin Amyloidosis. *Biochemical and Biophysical Research Communications* **2024**, *737*, 150770, doi:10.1016/j.bbrc.2024.150770.
11. Miyata, M.; Sato, T.; Kugimiya, M.; Sho, M.; Nakamura, T.; Ikemizu, S.; Chirifu, M.; Mizuguchi, M.; Nabeshima, Y.; Suwa, Y.; et al. The Crystal Structure of the Green Tea Polyphenol (−)-Epigallocatechin Gallate−Transthyretin Complex Reveals a Novel Binding Site Distinct from the Thyroxine Binding Site,. *Biochemistry* **2010**, *49*, 6104–6114, doi:10.1021/bi1004409.
12. Kristen, A.V.; Lehrke, S.; Buss, S.; Mereles, D.; Steen, H.; Ehlermann, P.; Hardt, S.; Giannitsis, E.; Schreiner, R.; Haberkorn, U.; et al. Green Tea Halts Progression of Cardiac Transthyretin Amyloidosis: An Observational Report. *Clin Res Cardiol* **2012**, *101*, 805–813, doi:10.1007/s00392-012-0463-z.
13. Powers, E.T.; Amass, L.; Baylor, L.; Fernández-Arias, I.; Riley, S.; Kelly, J.W. Transthyretin Kinetic Stabilizers for ATTR Amyloidosis: A Narrative Review of Mechanisms and Therapeutic Benefits. *Cardiol Ther* **2025**, *14*, 333–350, doi:10.1007/s40119-025-00423-7.
14. Eytcheson, S.A.; Zosel, A.D.; Olker, J.H.; Hornung, M.W.; Degitz, S.J. Screening the ToxCast Chemical Libraries for Binding to Transthyretin. *Chem. Res. Toxicol.* **2024**, *37*, 1670–1681, doi:10.1021/acs.chemrestox.4c00215.
15. Ciccone, L.; Tonali, N.; Nencetti, S.; Orlandini, E. Natural Compounds as Inhibitors of Transthyretin Amyloidosis and Neuroprotective Agents: Analysis of Structural Data for Future Drug Design. *Journal of Enzyme Inhibition and Medicinal Chemistry* **2020**, *35*, 1145–1162, doi:10.1080/14756366.2020.1760262.
16. Ueda, M.; Horibata, Y.; Shono, M.; Misumi, Y.; Oshima, T.; Su, Y.; Tasaki, M.; Shinriki, S.; Kawahara, S.; Jono, H.; et al. Clinicopathological Features of Senile Systemic Amyloidosis: An Ante- and Post-Mortem Study. *Modern Pathology* **2011**, *24*, 1533–1544, doi:10.1038/modpathol.2011.117.
17. Tetko, I.V. Tox24 Challenge. *Chem. Res. Toxicol.* **2024**, *37*, 825–826, doi:10.1021/acs.chemrestox.4c00192.
18. Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J Cheminform* **2018**, *10*, 4, doi:10.1186/s13321-018-0258-y.
19. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the Proceedings of the 31st International Conference on Neural Information Processing Systems; Curran Associates Inc.: Red Hook, NY, USA, December 2017; pp. 3149–3157.
20. Eytcheson, S.A.; Tetko, I.V. Which Modern AI Methods Provide Accurate Predictions of Toxicological End Points? Analysis of Tox24 Challenge Results. *Chem. Res. Toxicol.* **2025**, *38*, 1443–1451, doi:10.1021/acs.chemrestox.5c00273.
21. OECD Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models. *OECD Series on Testing and Assessment* **2014**, *No. 69*, doi:10.1787/9789264085442-en.
22. Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions 2017.
23. Klabunde, T.; Petrassi, H.M.; Oza, V.B.; Raman, P.; Kelly, J.W.; Sacchettini, J.C. Rational Design of Potent Human Transthyretin Amyloid Disease Inhibitors. *Nat Struct Biol* **2000**, *7*, 312–321, doi:10.1038/74082.
24. Oza, V.B.; Smith, C.; Raman, P.; Koepf, E.K.; Lashuel, H.A.; Petrassi, H.M.; Chiang, K.P.; Powers, E.T.; Sachettinni, J.; Kelly, J.W. Synthesis, Structure, and Activity of Diclofenac Analogues as Transthyretin Amyloid Fibril Formation Inhibitors. *J Med Chem* **2002**, *45*, 321–332, doi:10.1021/jm010257n.
25. Adamski-Werner, S.L.; Palaninathan, S.K.; Sacchettini, J.C.; Kelly, J.W. Diflunisal Analogues Stabilize the Native State of Transthyretin. Potent Inhibition of Amyloidogenesis. *J Med Chem* **2004**, *47*, 355–374, doi:10.1021/jm030347n.
26. Palaninathan, S.K.; Mohamedmohaideen, N.N.; Orlandini, E.; Ortore, G.; Nencetti, S.; Lapucci, A.; Rossello, A.; Freundlich, J.S.; Sacchettini, J.C. Novel Transthyretin Amyloid Fibril Formation Inhibitors: Synthesis, Biological Evaluation, and X-Ray Structural Analysis. *PLOS ONE* **2009**, *4*, e6290, doi:10.1371/journal.pone.0006290.
27. Institut National de la Recherche Agronomique. PhytoHub Available online: https://phytohub.eu/ (accessed on 23 December 2025).
28. Online Chemical Modeling Environment Available online: https://ochem.eu/static/challenge.do (accessed on 4 October 2024).

 doi:10.20944/preprints202604.1199.v1

29. Blake, C.C.F.; Geisow, M.J.; Oatley, S.J.; Rérat, B.; Rérat, C. Structure of Prealbumin: Secondary, Tertiary and Quaternary Interactions Determined by Fourier Refinement at 1.8 Å. *Journal of Molecular Biology* **1978**, *121*, 339–356, doi:10.1016/0022-2836(78)90368-6.
30. Trivella, D.B.B.; dos Reis, C.V.; Lima, L.M.T.R.; Foguel, D.; Polikarpov, I. Flavonoid Interactions with Human Transthyretin: Combined Structural and Thermodynamic Analysis. *Journal of Structural Biology* **2012**, *180*, 143–153, doi:10.1016/j.jsb.2012.07.008.
31. Florio, P.; Folli, C.; Cianci, M.; Rio, D.D.; Zanotti, G.; Berni, R. Transthyretin Binding Heterogeneity and Anti-Amyloidogenic Activity of Natural Polyphenols and Their Metabolites*. *Journal of Biological Chemistry* **2015**, *290*, 29769–29780, doi:10.1074/jbc.M115.690172.
32. Yokoyama, T.; Ueda, M.; Ando, Y.; Mizuguchi, M. Discovery of γ-Mangostin as an Amyloidogenesis Inhibitor. *Sci Rep* **2015**, *5*, 13570, doi:10.1038/srep13570.
33. Iakovleva, I.; Begum, A.; Pokrzywa, M.; Walfridsson, M.; Sauer-Eriksson, A.E.; Olofsson, A. The Flavonoid Luteolin, but Not Luteolin-7-O-Glucoside, Prevents a Transthyretin Mediated Toxic Response. *PLOS ONE* **2015**, *10*, e0128222, doi:10.1371/journal.pone.0128222.
34. Ferreira, N.; Saraiva, M.J.; Almeida, M.R. Natural Polyphenols Inhibit Different Steps of the Process of Transthyretin (TTR) Amyloid Fibril Formation. *FEBS Letters* **2011**, *585*, 2424–2430, doi:10.1016/j.febslet.2011.06.030.
35. Pullakhandam, R.; Srinivas, P.N.B.S.; Nair, M.K.; Reddy, G.B. Binding and Stabilization of Transthyretin by Curcumin. *Archives of Biochemistry and Biophysics* **2009**, *485*, 115–119, doi:10.1016/j.abb.2009.02.013.
36. Griswold, M.G.; Fullman, N.; Hawley, C.; Arian, N.; Zimsen, S.R.M.; Tymeson, H.D.; Venkateswaran, V.; Tapp, A.D.; Forouzanfar, M.H.; Salama, J.S.; et al. Alcohol Use and Burden for 195 Countries and Territories, 1990–2016: A Systematic Analysis for the Global Burden of Disease Study 2016. *The Lancet* **2018**, *392*, 1015–1035, doi:10.1016/S0140-6736(18)31310-2.
37. Maggiora, G.M. On Outliers and Activity CliffsWhy QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535–1535, doi:10.1021/ci060117s.
38. Gade, A.; Kumar, M.S. Gut Microbial Metabolites of Dietary Polyphenols and Their Potential Role in Human Health and Diseases. *J Physiol Biochem* **2023**, *79*, 695–718, doi:10.1007/s13105-023-00981-1.