Cell Reports Medicine, Volume *5*

Supplemental information

Machine learning-based analysis identifies and validates serum exosomal proteomic signatures for the diagnosis of colorectal cancer

Haofan Yin, Jinye Xie, Shan Xing, Xiaofang Lu, Yu Yu, Yong Ren, Jian Tao, Guirong He, Lijun Zhang, Xiaopeng Yuan, Zheng Yang, and Zhijian Huang

Supplementary materials

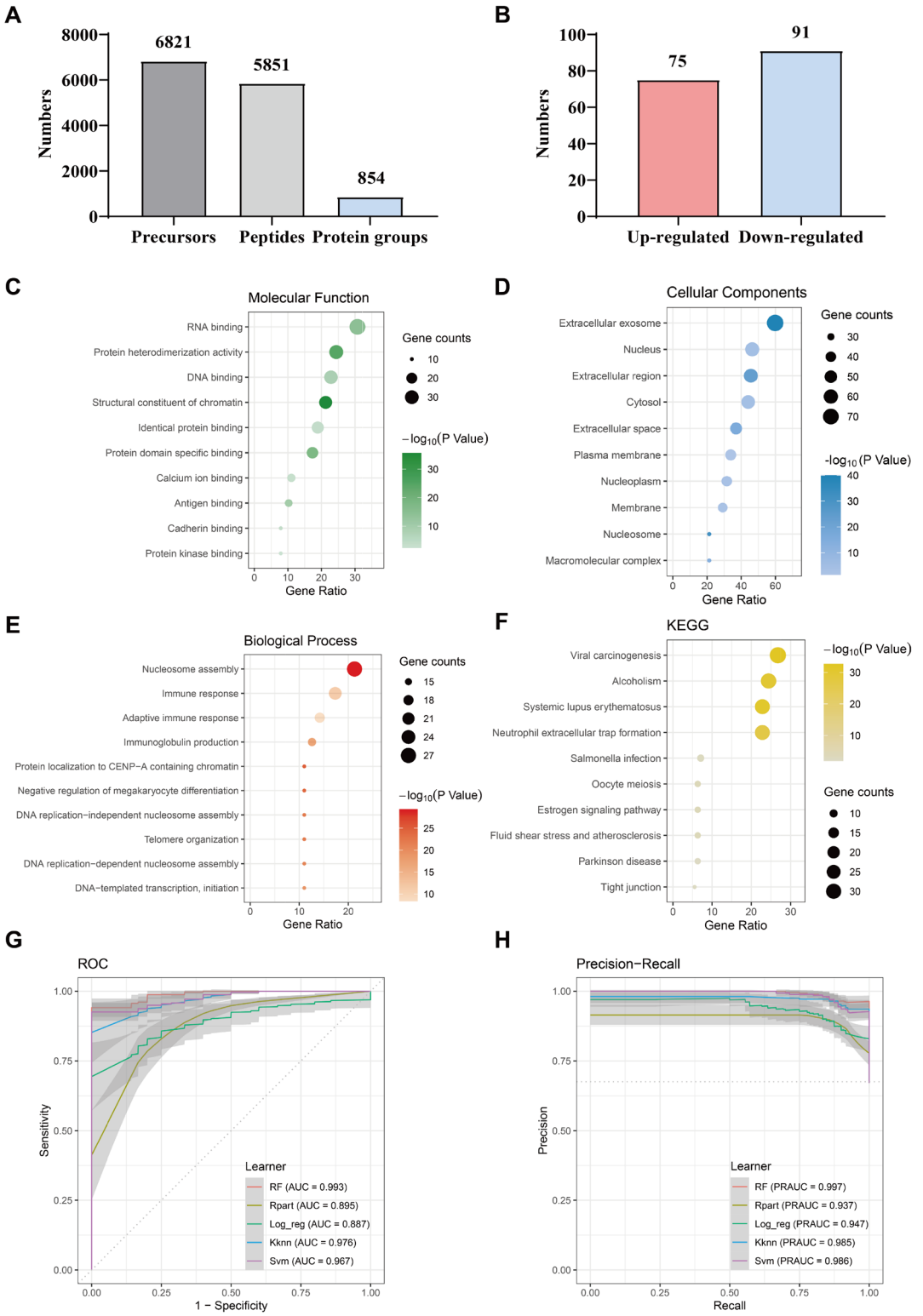

Figure S1 DEPs identification and functional enrichment analysis in serum EVs based on 4D-DIA proteomics.

Number of precursors, peptides, and proteins from serum EVs identified by 4D-DIA proteomics in CRC patients and HC. (B) Number of up-regulated and down-regulated proteins in serum EVs of CRC. (C-F) Down-regulated proteins enrichment analysis

displayed the potential molecular function (C), cellular component (D), biological process (E), and KEGG pathways (F) enriched in HC compared to CRC group. (G, H) Diagnostic performance of ML models based on different algorithms. ROC (G) and PR (H) curves of the indicated ML diagnostic models based on the 12 candidate proteins.

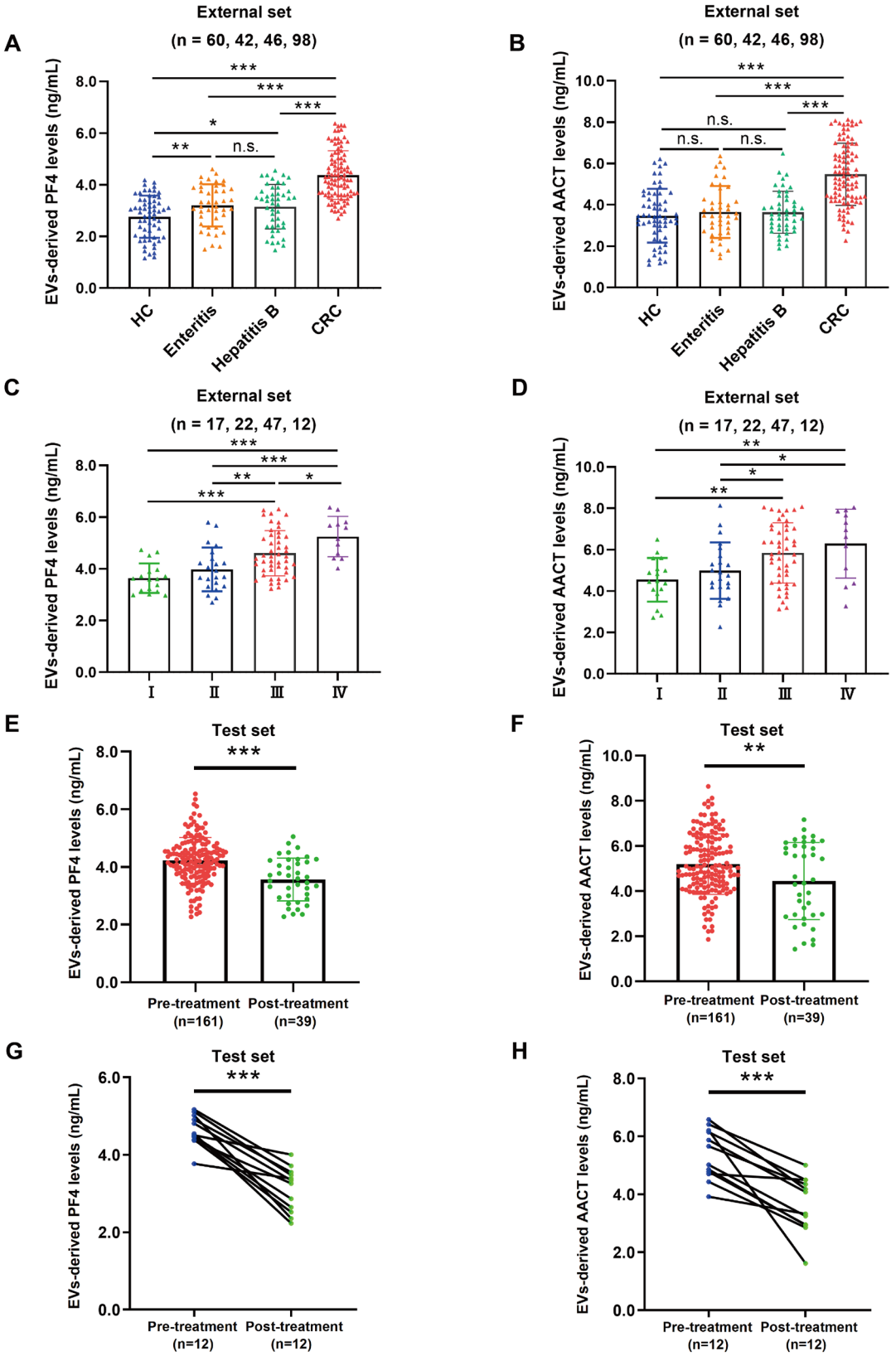

Figure S2 Differential levels of PF4 and AACT derived from EVs among different clinical groups.

(A, B) EVs-derived PF4 (A) and AACT (B) levels detected by ELISA in HC (n = 60), Enteritis (n = 42), Hepatitis B (n = 46), and CRC (n = 98) groups from the external set. (C, D) The levels of EVs-derived PF4 (C) and AACT (D) at different clinical stages (I: n = 17, II: n = 22, III: n = 47, IV:n = 12) of CRC patients in the external set. (E, F) Comparison of EVs-derived PF4 (E) and AACT (F) levels in CRC patients with (n=39) and without (n=161) treatment in the train set. (G, H) Comparison of EVs-derived PF4 (G) and AACT (H) levels before and after treatment in 12 CRC patients. Data are shown as mean ± SD; n.s.: Not significance, * P <0.05, ** P <0.01, and *** P <0.001.

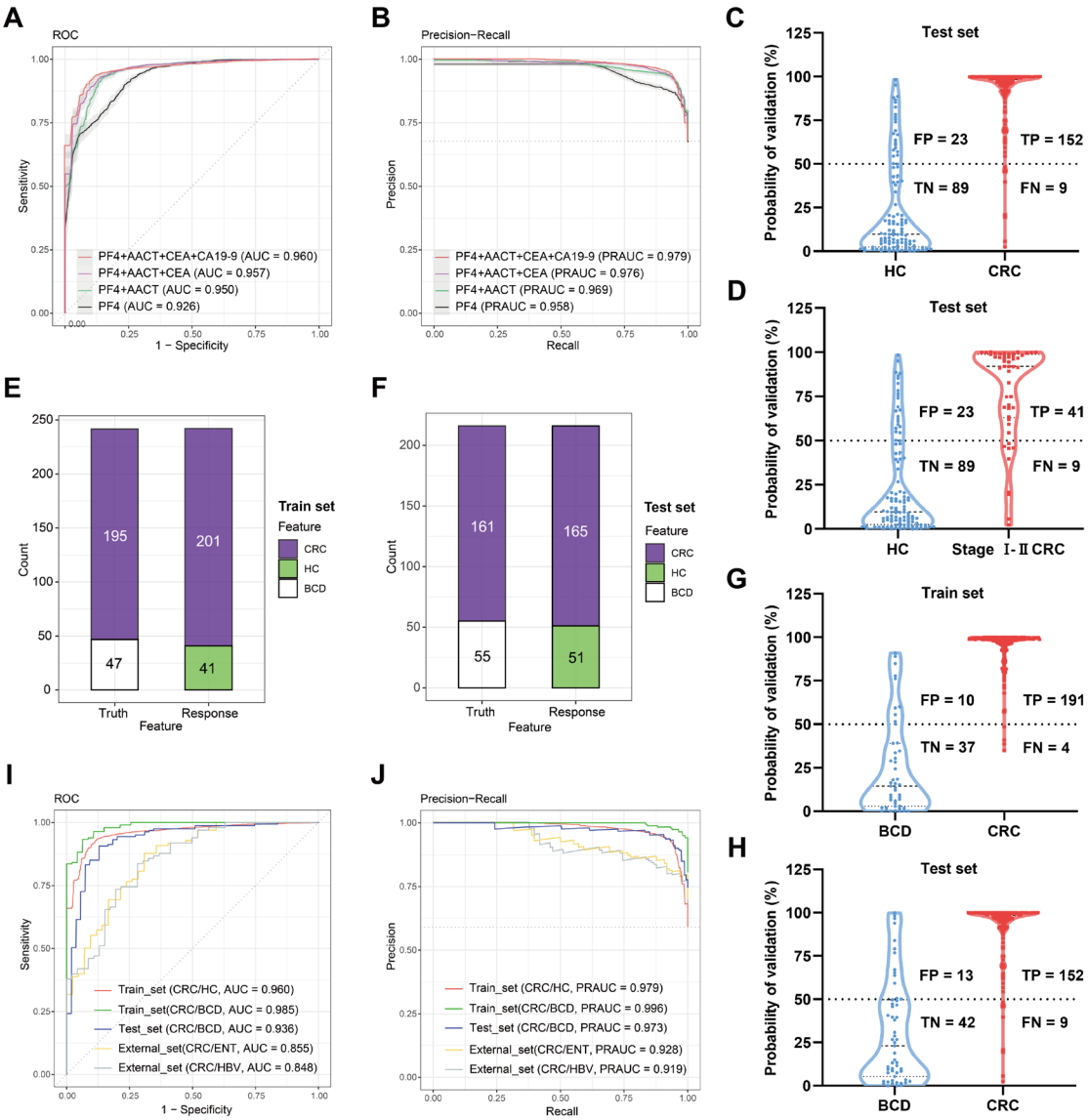

Figure S3 Diagnostic performance of different combinations and the EVs-related model for distinguishing CRC from BCD and inflammatory disease.

(A, B) ROC (A) and PR (B) curves of the indicated ML diagnostic models based on different combinations of 4 variables. (C) Probability of responding CRC in validating the EVs-related model using the test set. (D) Probability of responding CRC in

validating the EVs-related model using stage I-II and HC samples from the test set. (E, F) Confusion matrix displayed the prediction results for 242 train set (E) (CRC: 195 cases, BCD: 47 cases) and 216 test set (F) (CRC: 161 cases, BCD: 55 cases) through the EVs-related diagnostic model. (G, H) Probability of responding CRC in validating the EVs-related model using CRC and BCD samples from the train (G) and test (H) sets. (I, J) ROC curve (I) and PR curve (J) were plotted for the EVs-related diagnostic model using the train, test, and external sets.

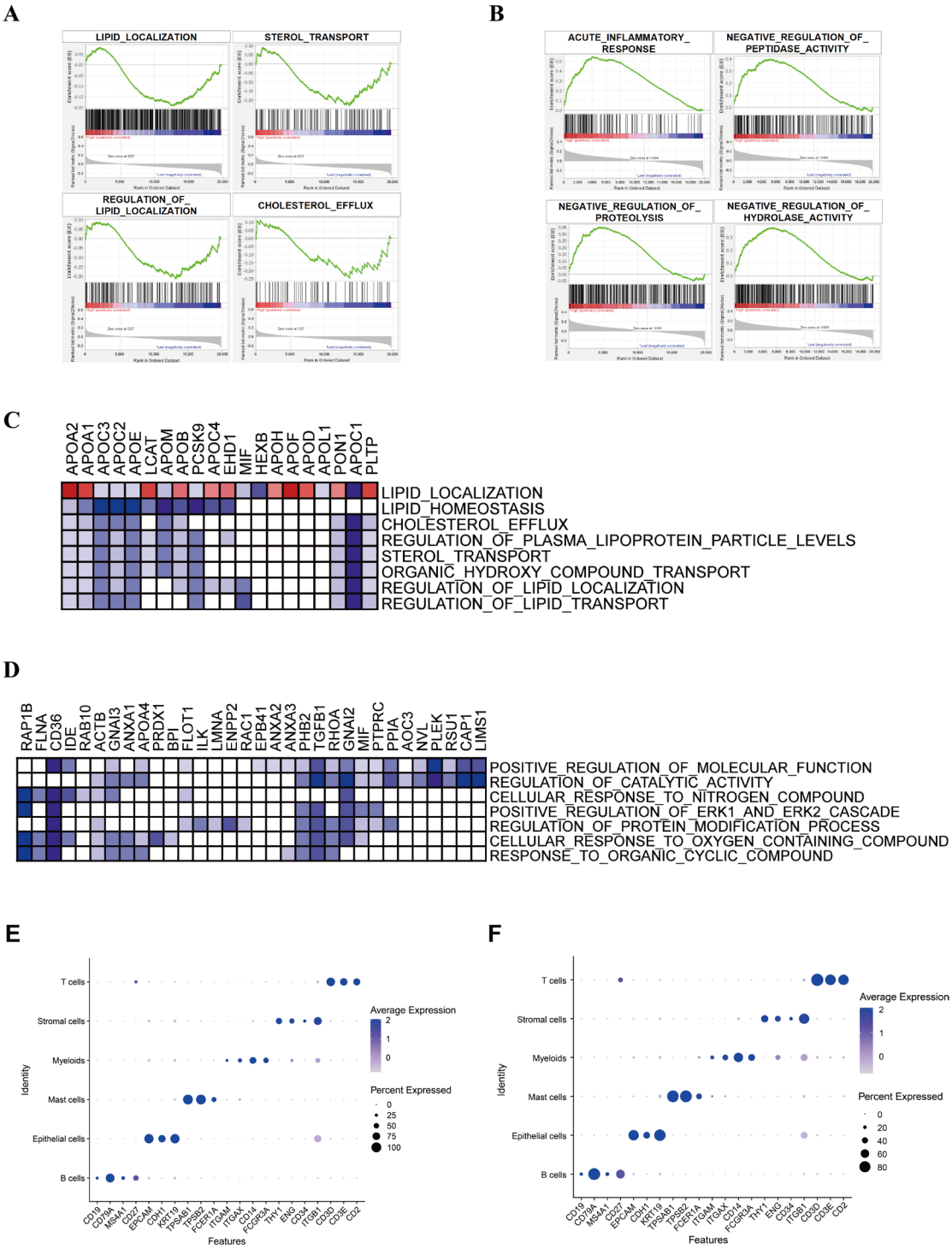

Figure S4 Functional enrichment analysis of PF4 and AACT.

(A, B) GSEA illustrated the pathways enriched in PF4-low phenotype (A) and AACT-high phenotype (B) in TCGA CRC database. (C, D) Leading edge analysis of the intersection proteins in the EnrichmentMap pathways network of PF4-low (C) and AACT-low (D) phenotypes. (E, F) Dotplot of cell markers in various cell types. Dotplot showed the expression of cell markers in various cell types from the GSE132465 dataset (E) and the GSE132257 dataset (F).

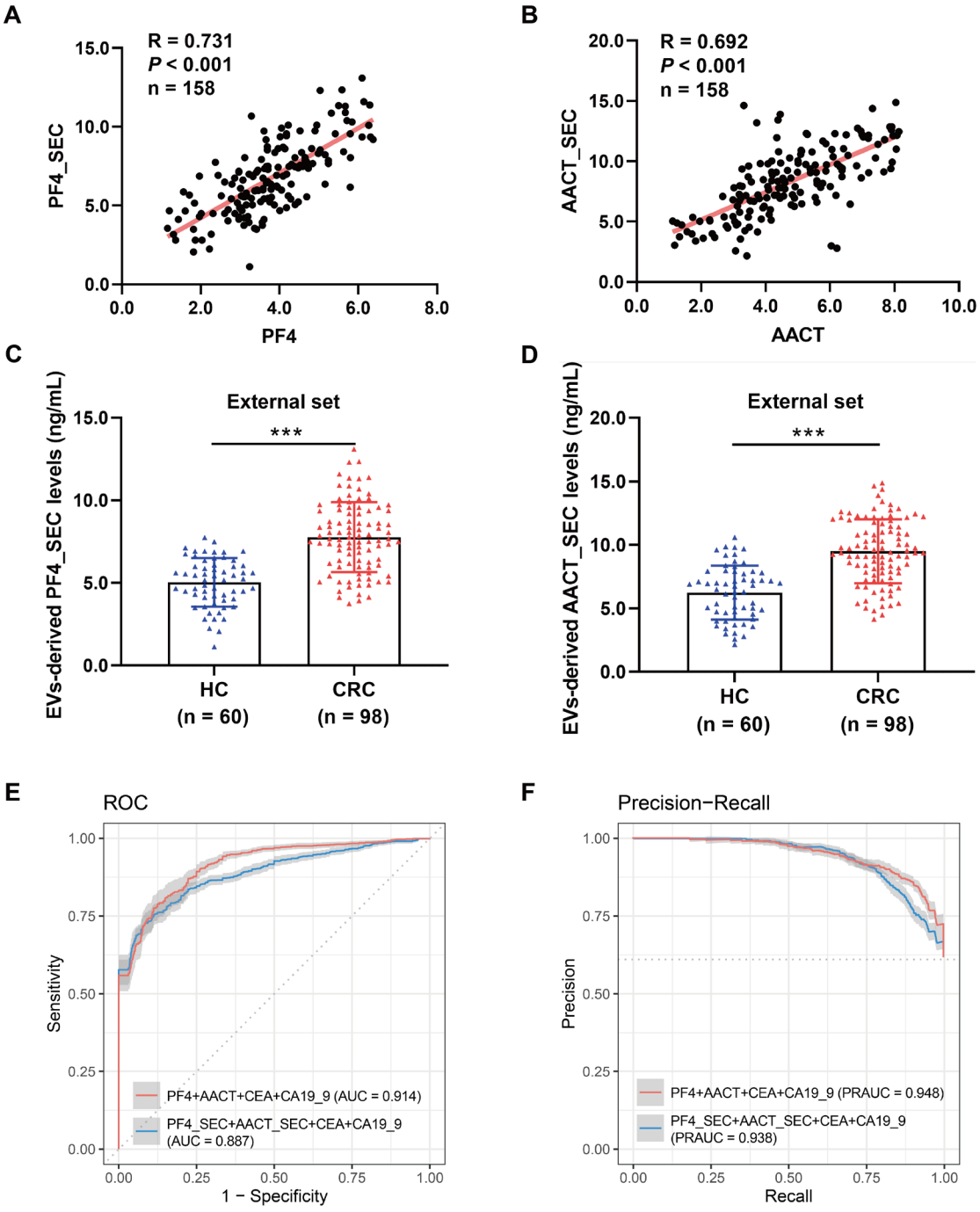

Figure S5 Diagnostic assessment of PF4 and AACT extracted by size exclusion chromatography for CRC.

(A, B) Pearson correlation analysis of PF4 (A) and AACT (B) isolated by

ultracentrifugation and SEC methods. n = 158. (C, D) EVs-derived PF4 (C) and AACT (D) isolated by size exclusion chromatography (SEC) method were detected by ELISA in HC (n = 60) and CRC (n = 98) groups from the external set. (E) ROC (E) and PR curve (F) of RF diagnostic models based on SEC isolated PF4, SEC isolated AACT, CEA, and CA19_9.

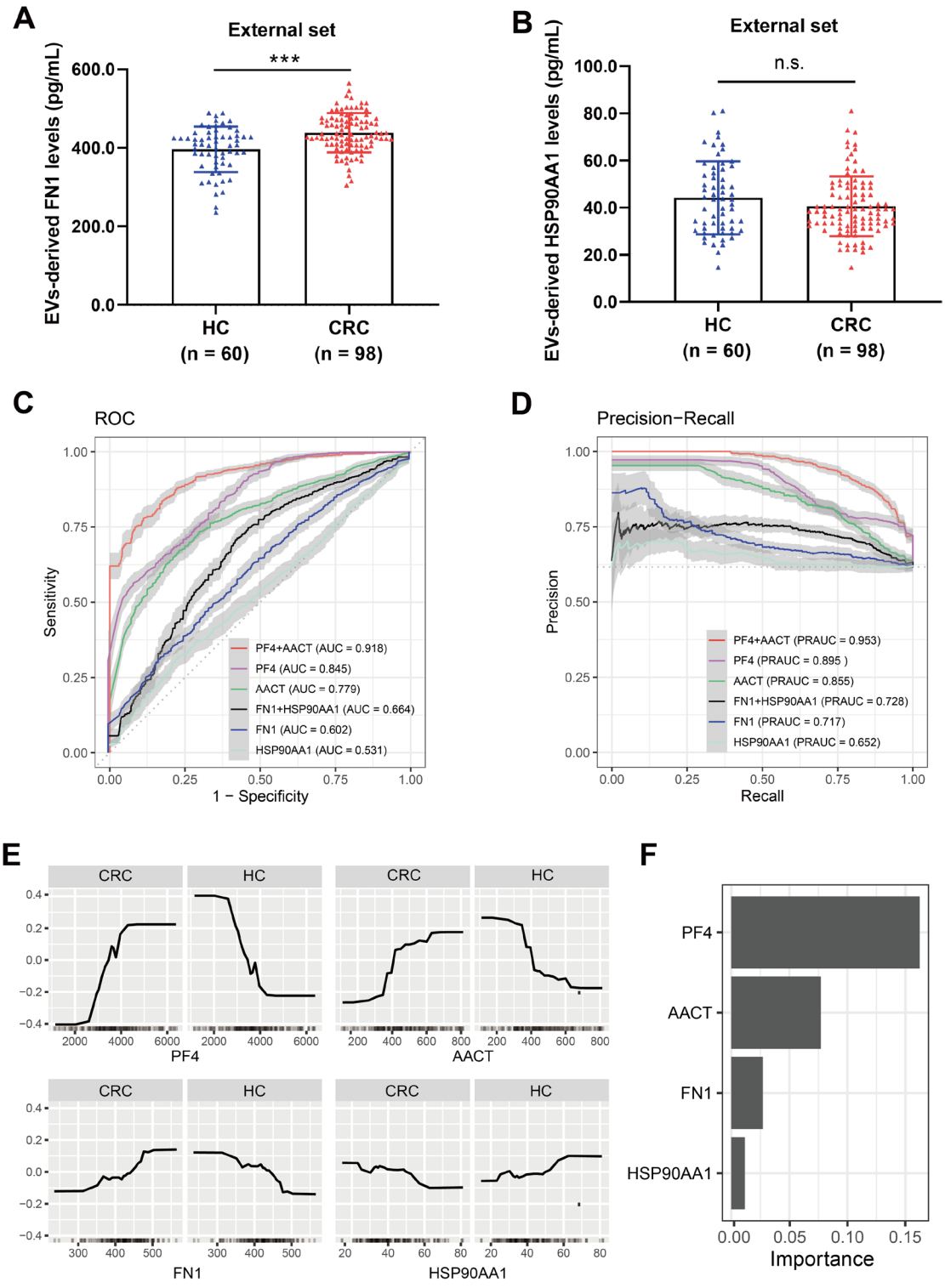

Figure S6 Diagnostic performance of other biomarkers proposed in competitive

study.

(A, B) EVs-derived FN1 (A) and HSP90AA1 (B) levels detected by ELISA in HC (n = 60) and CRC (n = 98) groups from the external set. (C, D) ROC curve (C) and PR curve (D) of RF diagnostic models based on indicated variables in the external set. (E) The ALE curve depicts the accumulated local effects of PF4, AACT, FN1, and HSP90AA1. The x-axis represents the feature values, and the y-axis represents the accumulated local effects. (F) Variable importance score plot showed the contribution of 4 variables in the RF diagnostic model.

Table S1. Performance of different ML models.

ML Algorithms	AUC	PRAUC	Classification error	Sensitivity	Specificity	Precision	Recall	Accuracy	F1 score
RF	0.993	0.997	0.077	0.979	0.804	0.918	0.979	0.923	0.948
Rpart	0.895	0.937	0.083	0.939	0.852	0.948	0.939	0.917	0.943
Log_reg	0.887	0.947	0.196	0.800	0.832	0.913	0.800	0.804	0.852
Kknn	0.976	0.985	0.070	0.986	0.802	0.923	0.986	0.930	0.954
SVM	0.967	0.986	0.117	0.966	0.705	0.880	0.966	0.883	0.921

Table S2. Venn result of RF model variables and LASSO model variables.

Names	Total	Elements
RF_5_variables & LASSO_λ_1se & LASSO_λ_min	2	PF4; AACT
LASSO _λ_1se LASSO _λ_min	6	PGRP1; PGBM; X1433B; CALL3; S10A6; MA1A1
LASSO _λ_min	7	PCSK6; HV205; IGJ; TGFB1; RETN; ATPB; FA11
LASSO _λ_1se	2	AJM1; IF5A1_AND_IF5AL
RF_5_variables	3	KLD8B; CP135; KV315

Table S3. Correlation between serum EVs-derived PF4 and clinicopathological characteristics of CRC patients in the train set, test set, and external set.

Characteristics	Train Set (n=195)			Test Set (n=161)			External Set (n=98)		
	No.of patients	PF4 level (mean±SD)	*P* value	No.of patients	PF4 level (mean±SD)	*P* value	No.of patients	PF4 level (mean±SD)	*P* value
Age									
⩽60	103	4.079±0.824	0.674	82	4.269±0.855	0.544	38	4.384±1.039	0.928
>60	92	4.127±0.784		79	4.193±0.730		60	4.366±0.887	
Gender									
Male	111	4.105±0.783	0.985	98	4.153±0.786	0.116	47	4.287±0.955	0.382
Female	84	4.103±0.830		63	4.355±0.799		51	4.452±0.936	
Clinical stage									
I-II	70	3.624±0.783	<0.001	50	3.658±0.687	<0.001	39	3.829±0.749	<0.001
III-IV	125	4.373±0.747		111	4.490±0.701		59	4.732±0.889	
T classification									
T1-T2	49	3.822±0.789	0.003	32	3.798±1.034	0.001	20	3.826±0.775	0.003

T3-T4	146	4.201±0.785		129	4.335±0.692		77	4.515±0.942	
N classification									
N0	73	3.682±0.713	<0.001	55	3.741±0.724	<0.001	39	3.829±0.749	<0.001
N1-N3	122	4.357±0.745		106	4.487±0.708		58	4.738±0.896	
M classification									
M0	153	3.975±0.703	<0.001	97	3.931±0.696	<0.001	85	4.254±0.907	0.001
M1	42	4.575±0.956		64	4.688±0.718		13	5.147±0.827	
Differentiation									
Poor	24	3.956±0.828	0.351	25	4.561±0.657	0.025	15	4.615±0.954	0.259
Moderate / Well	111	4.139±0.879		120	4.163±0.825		64	4.298±0.977	
HER2 expression									
Negative	88	4.070±0.806	0.459	64	4.180±0.752	0.768	36	4.289±1.005	0.725
Positive	36	4.191±0.877		38	4.223±0.624		23	4.381±0.920	
BRAF status									
Wild	113	4.108±0.844	0.590	43	4.386±0.585	0.016	53	4.362±0.962	NA
Mutation	6	3.918±0.733		4	5.148±0.525		0	NA	
MSS/MSI status									
MSS	124	4.173±0.866	0.114	107	4.207±0.736	0.390	61	4.282±1.044	0.397
MSI	11	3.749±0.546		9	3.98±0.85		7	4.626±0.618	
CEA (ng/mL)									
<5	129	4.031±0.761	0.073	90	4.056±0.752	0.002	65	4.256±0.909	0.087
≥5	66	4.248±0.863		71	4.446±0.797		33	4.602±0.981	
CA19-9 (ng/mL)									
<35	153	4.064±0.779	0.184	112	4.093±0.726	0.001	82	4.353±0.942	0.638
≥35	42	4.250±0.873		49	4.558±0.866		16	4.475±0.977	

Table S4. Correlation between serum EVs-derived AACT and clinicopathological characteristics of CRC patients in the train set, test set, and external set.

Characteristics	Train Set (n=195)			Test Set (n=161)			External Set (n=98)		
	No.of patients	AACT level (mean±SD)	*P* value	No.of patients	AACT level (mean±SD)	*P* value	No.of patients	AACT level (mean±SD)	*P* value
Age									
⩽60	103	5.204±1.422	0.306	82	5.172±1.426	0.807	38	5.703±1.528	0.235
>60	92	5.409±1.357		79	5.224±1.253		60	5.331±1.481	
Gender,									
Male	111	5.303±1.414	0.971	98	5.157±1.411	0.638	47	5.325±1.554	0.345
Female	84	5.297±1.370		63	5.260±1.229		51	5.614±1.456	
Clinical stage									

I-II	70	4.923±1.393	0.004	50	4.681±1.287	0.001	39	4.791±1.244	<0.001
III-IV	125	5.512±1.351		111	5.430±1.303		59	5.928±1.496	
T classification									
T1-T2	49	5.013±1.331	0.095	32	4.777±1.293	0.051	20	4.991±1.466	0.122
T3-T4	146	5.397±1.403		129	5.298±1.336		77	5.568±1.479	
N classification									
N0	73	4.910±1.400	0.002	55	4.733±1.264	0.001	39	4.791±1.244	<0.001
N1-N3	122	5.534±1.339		106	5.439±1.320		58	5.892±1.483	
M classification									
M0	153	5.159±1.347	0.006	97	4.983±1.389	0.012	85	5.373±1.459	0.084
M1	42	5.816±1.446		64	5.523±1.201		13	6.147±1.669	
Differentiation									
Poor	24	5.704±1.630	0.137	25	5.277±1.351	0.793	15	5.479±1.500	0.880
Moderate / Well	111	5.236±1.333		120	5.200±1.318		64	5.546±1.542	
HER2 expression									
Negative	88	5.443±1.407	0.259	64	5.241±1.299	0.377	36	5.684±1.662	0.393
Positive	36	5.120±1.515		38	4.997±1.407		23	5.331±1.298	
BRAF status									
Wild	113	5.400±1.463	0.241	43	5.282±1.104	0.849	53	5.645±1.583	NA
Mutation	6	4.678±1.449		4	5.172±1.020		0	NA	
MSS/MSI status									
MSS	124	5.384±1.387	0.818	107	5.025±1.258	0.711	61	5.477±1.535	0.838
MSI	11	5.283±1.496		9	5.190±1.286		7	5.604±1.731	
CEA (ng/mL)									
<5	129	5.178±1.393	0.086	90	5.056±1.346	0.166	65	5.412±1.555	0.561
≥5	66	5.540±1.369		71	5.352±1.312		33	5.600±1.408	
CA19-9 (ng/mL)									
<35	153	5.220±1.420	0.123	112	5.121.±1.401	0.193	82	5.417±1.429	0.388
≥35	42	5.595±1.258		49	5.421±1.144		16	5.774±1.859	

Table S5. Diagnostic performance of RF models based on different combinations.

Combinations	AUC	PRAUC	Classification error	Sensitivity	Specificity	Precision	Recall	Accuracy	F1 score
PF4	0.926	0.958	0.146	0.894	0.771	0.891	0.894	0.854	0.893
PF4+AACT	0.950	0.969	0.089	0.945	0.843	0.924	0.945	0.911	0.935
PF4+AACT+CEA	0.957	0.976	0.092	0.937	0.849	0.927	0.937	0.908	0.932
PF4+AACT+ CEA+CA199	0.960	0.979	0.084	0.942	0.867	0.935	0.942	0.916	0.938

Table S6. Diagnostic performances of train set, test set, and external set through the EVs-related model.

Models	AUC	PRAUC	Classification error	Sensitivity	Specificity	Precision	Recall	Accuracy	F1 score
The EVs-related model	0.960	0.979	0.084	0.942	0.867	0.935	0.942	0.916	0.938
Test set (CRC vs HC)	0.963	0.975	0.117	0.944	0.795	0.869	0.944	0.883	0.905
Test set (Stage I-II CRC vs HC)	0.905	0.841	0.198	0.820	0.795	0.641	0.820	0.802	0.719
Train set (CRC vs BCD)	0.985	0.996	0.058	0.979	0.787	0.950	0.979	0.942	0.965
Test set (CRC vs BCD)	0.936	0.973	0.102	0.944	0.764	0.921	0.944	0.898	0.933
External set (CRC vs HC)	0.895	0.921	0.190	0.959	0.567	0.783	0.959	0.810	0.862
External set (CRC vs Enteritis)	0.855	0.928	0.186	0.959	0.476	0.810	0.959	0.814	0.879
External set (CRC vs Hepatitis B)	0.848	0.919	0.194	0.959	0.478	0.797	0.959	0.806	0.870