Supplemental information

ACP-TransLSTM: A Novel Deep Learning Framework for Anticancer Peptide Prediction Using Multi-Source Feature Integration

Jinxin Liu[1], Zhenming Wu[1] and Jin Zhao[1*]

[1] School of Computer Science and Technology, Qingdao University, Ningxia Road, 266071, Qingdao, Shandong, China
zhaojin@qdu.edu.cn

Contents

1. Supplemental Notes

Supplemental Note 1.1: Hyperparameter Setting

In the CNN + Bi-LSTM model, we first tested the CNNs individually and in order to find the best hidden layer settings, we chose different numbers of filter layers. We selected six different filter sizes: 8, 16, 32, 64, 128 and 256. As illustrated in Table S16, the model with 256 filters achieved the best performance and the highest AUC value. The choice of 256 filters is a trade-off between computational complexity and the ability to capture diverse local patterns in peptide sequences. A larger number of filters can potentially capture more complex patterns but may lead to overfitting and increased computational cost. When we decreased the number of CNN filters from 256 to 128, the average ACC across the six datasets decreased from 0.923 to 0.918, and the AUC decreased from 0.984 to 0.976. This indicates that reducing the number of filters weakens the model's ability to capture local sequence patterns, leading to a decline in performance. Therefore, we chose to use 256 filters when building the model. After determining the parameters of the CNN model, we compared the performance of different hidden cell counts (8, 16, 32, 64, 128, 256) in order to find the optimal hidden cell settings. The model with 64 hidden cells achieving the best performance and the highest AUC value. This number of units is determined through a series of preliminary experiments. A smaller number of units may not be able to fully capture the long-range dependencies in peptide sequences, resulting in suboptimal performance. Conversely, an excessive number of units can cause overfitting. Adjusting the number of Bi-LSTM units also affected the model's performance. When the number of units was reduced to 32, as shown in Table S17, the ACC across the six datasets dropped from 0.916 to 0.902, and the MCC decreased from 0.88 to 0.856. This shows that fewer units limit the model's ability to capture long-range dependencies. On the other hand, increasing the number of units to 128 led to overfitting.

To accurately determine the optimal critical probability for differentiating between ACPs and non-ACPs, we carried out a series of experiments on six datasets. We selected several different probability values, namely 0.3, 0.4, 0.5, 0.6 and 0.7 for testing. The experimental results show that when the critical probability is set to 0.5, the model exhibits the most excellent performance.

2. Supplemental Figures

Supplemental Figure 2.1: Amino acid composition information

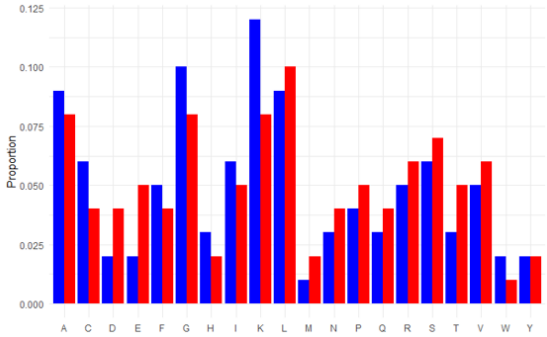

Fig. S1. Amino acid composition of ACPs and non-ACPs on ACP240 dataset. The type represented by blue is ACPs, and the type represented by red is non-ACPs.

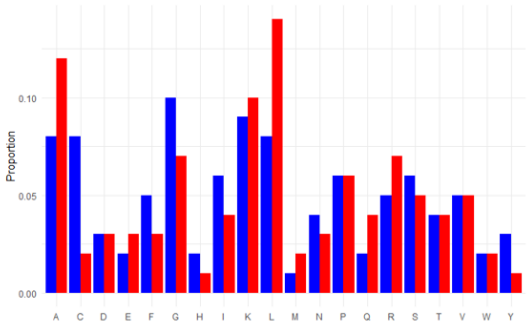

Fig. S2. Amino acid composition of ACPs and non-ACPs on ACP740 dataset. The type represented by blue is ACPs, and the type represented by red is non-ACPs.

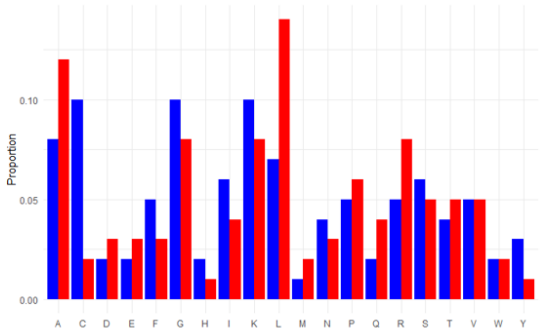

Fig. S3. Amino acid composition of ACPs and non-ACPs on ACP530 dataset. The type represented by blue is ACPs, and the type represented by red is non-ACPs.

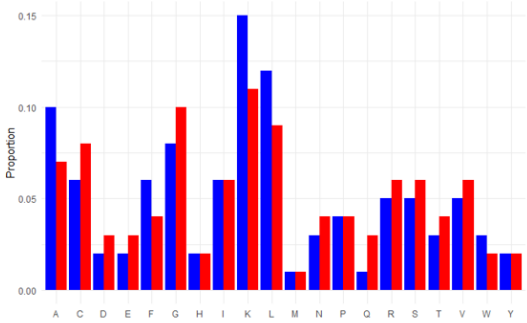

Fig. S4. Amino acid composition of ACPs and non-ACPs on ACPmain dataset. The type represented by blue is ACPs, and the type represented by red is non-ACPs.

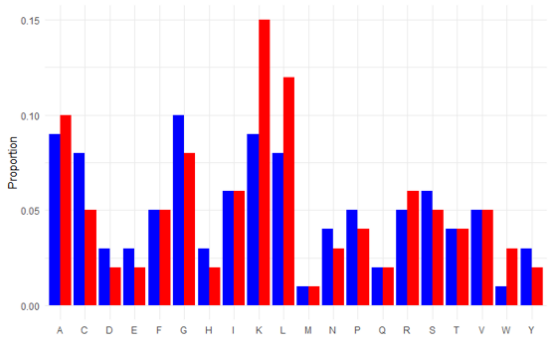

Fig. S5. Amino acid composition of ACPs and non-ACPs on ACPred-FL dataset. The type represented by blue is ACPs, and the type represented by red is non-ACPs.

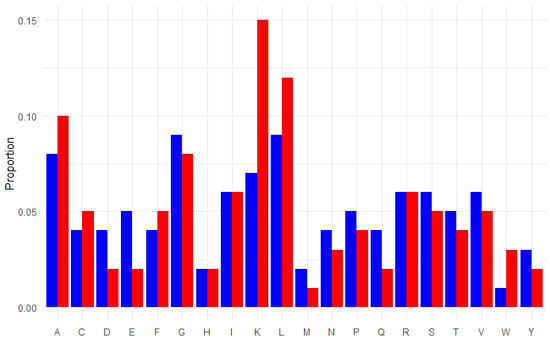

Fig. S6. Amino acid composition of ACPs and non-ACPs on ACPred-Fuse dataset. The type represented by blue is ACPs, and the type represented by red is non-ACPs.

Supplemental Figure 2.2: Charge distribution

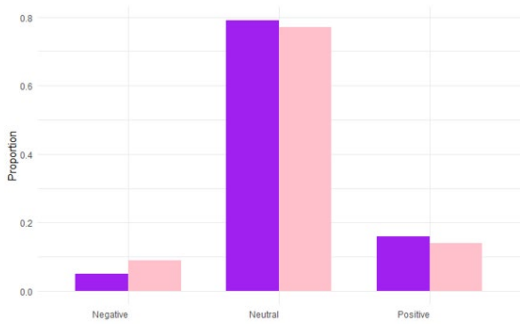

Fig. S7. Charge distribution of ACPs and non-ACPs on ACP240 dataset. The type represented by purple is ACPs, and the type represented by pink is non-ACPs.

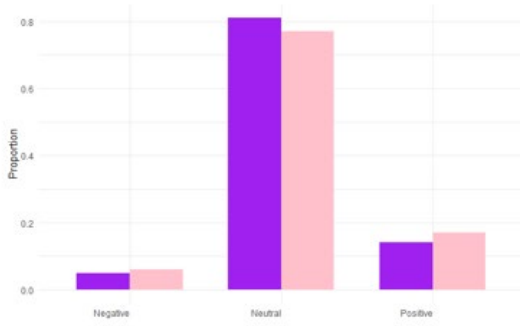

Fig. S8. Charge distribution of ACPs and non-ACPs on ACP740 dataset. The type represented by purple is ACPs, and the type represented by pink is non-ACPs.

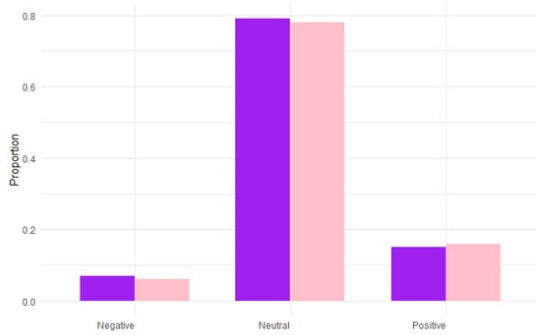

Fig. S9. Charge distribution of ACPs and non-ACPs on ACP530 dataset. The type represented by purple is ACPs, and the type represented by pink is non-ACPs.

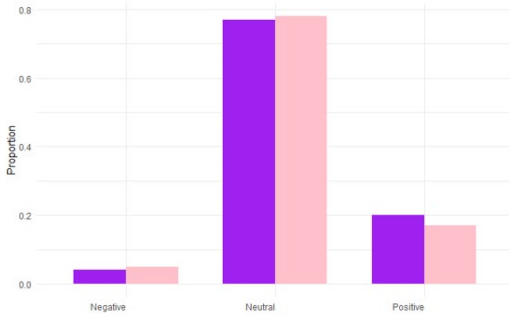

Fig. S10. Charge distribution of ACPs and non-ACPs on ACPmain dataset. The type represented by purple is ACPs, and the type represented by pink is non-ACPs.

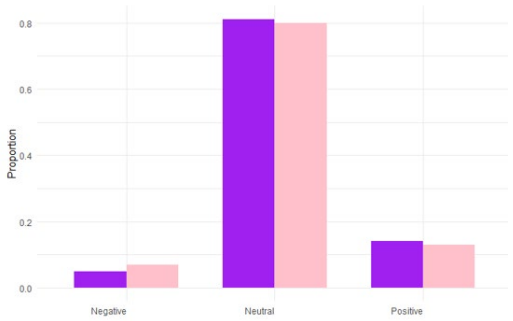

Fig. Fig. S11. Charge distribution of ACPs and non-ACPs on ACPred-FL dataset. The type represented by purple is ACPs, and the type represented by pink is non-ACPs.

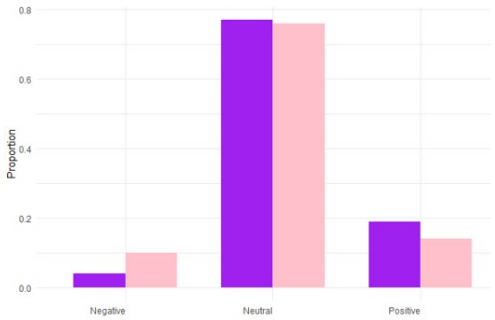

Fig. S11. Charge distribution of ACPs and non-ACPs on ACPred-Fuse dataset. The type represented by purple is ACPs, and the type represented by pink is non-ACPs.

Supplemental Figure 2.3: Length distribution

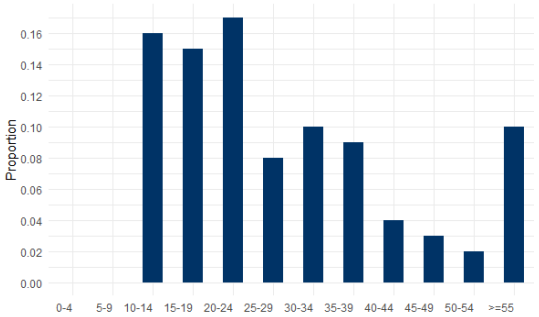

Fig. S12. Length distribution of ACPs and non-ACPs on ACP240 dataset.

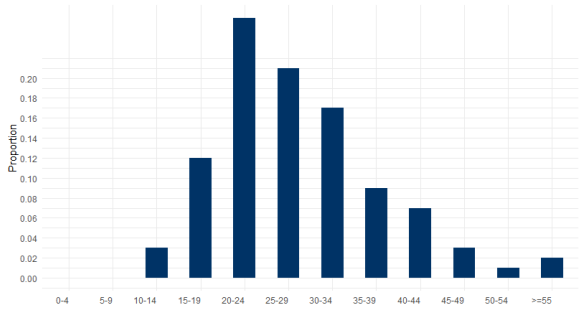

Fig. S13. Length distribution of ACPs and non-ACPs on ACP740 dataset.

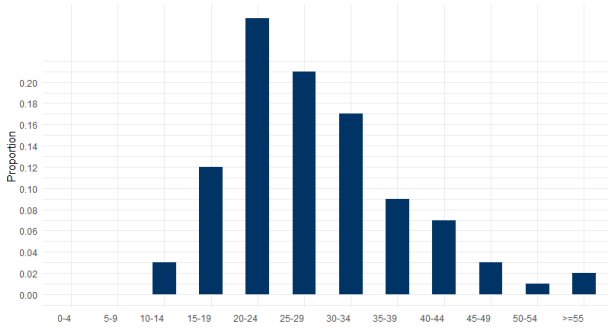

Fig. S14. Length distribution of ACPs and non-ACPs on ACP530 dataset.

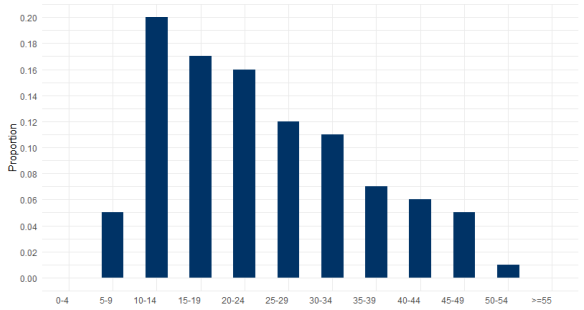

Fig. S15. Length distribution of ACPs and non-ACPs on ACPmain dataset.

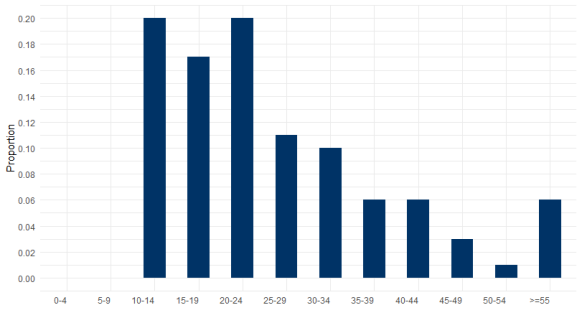

Fig. S16. Length distribution of ACPs and non-ACPs on ACPred-FL dataset.

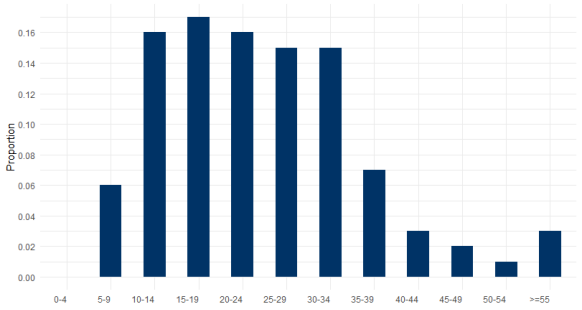

Fig. S17. Length distribution of ACPs and non-ACPs on ACPred-Fuse dataset.

3. Supplemental Tables

Supplemental Table 3.1: Experimental Results of different filters

Table S1. Performance comparison of different filters of CNN on ACP530 dataset.

Dataset	Filter	ACC	SE	SP	F1	MCC	AUC
ACP530	8	0.915	0.926	0.905	0.916	0.834	0.979
	16	0.925	0.925	0.921	0.925	0.853	0.98
	32	0.91	0.902	0.918	0.908	0.828	0.981
	64	0.924	**0.936**	0.911	0.925	0.851	**0.984**
	128	0.834	0.785	0.883	0.816	0.7	0.963
	256	**0.931**	0.933	**0.929**	**0.931**	**0.863**	**0.984**

Table S2. Performance comparison of different filters of CNN on ACP240 dataset.

Dataset	Filter	ACC	SE	SP	F1	MCC	AUC
ACP240	8	0.912	0.903	0.921	0.911	0.824	0.972
	16	0.915	0.939	0.89	0.917	0.831	0.975
	32	0.928	0.907	0.949	0.927	0.857	0.982
	64	0.924	0.936	0.911	0.925	0.851	**0.984**
	128	0.929	**0.946**	0.912	0.931	0.86	0.981
	256	**0.931**	0.911	**0.951**	**0.932**	**0.863**	0.982

Table S3. Performance comparison of different filters of CNN on ACP740 dataset.

Dataset	Filter	ACC	SE	SP	F1	MCC	AUC
ACP740	8	0.919	0.947	0.891	0.921	0.84	0.977
	16	0.918	0.9	0.934	0.917	0.837	0.975
	32	0.92	0.889	**0.952**	0.918	0.843	0.979
	64	0.921	**0.954**	0.888	0.924	0.845	0.979
	128	0.909	0.959	0.858	0.913	0.822	0.975
	256	**0.927**	0.921	0.932	**0.926**	**0.854**	**0.98**

Table S4. Performance comparison of different filters of CNN on ACPmain dataset.

Dataset	Filter	ACC	SE	SP	F1	MCC	AUC
ACPmain	8	0.913	0.942	0.884	0.916	0.828	0.974
	16	0.918	0.915	0.922	0.918	0.837	0.976
	32	0.919	0.95	0.888	0.922	0.84	0.978
	64	**0.923**	0.923	**0.923**	0.923	0.847	0.978
	128	0.918	0.939	0.897	0.92	0.838	0.976
	256	**0.923**	**0.972**	0.874	**0.927**	**0.851**	**0.984**

Table S5. Performance comparison of different filters of CNN on ACPred-Fuse dataset.

Dataset	Filter	ACC	SE	SP	F1	MCC	AUC
ACPred-Fuse	8	0.914	0.916	0.912	0.915	0.83	0.974
	16	0.923	0.887	**0.959**	0.92	0.849	0.981
	32	**0.925**	0.914	0.935	0.924	**0.851**	0.98
	64	0.922	0.911	0.933	0.92	0.84	0.978
	128	0.92	0.9	0.933	0.918	0.84	0.976
	256	0.924	**0.959**	0.889	**0.926**	**0.851**	**0.983**

Table S6. Performance comparison of different filters of CNN on ACPred-FL dataset.

Dataset	Filter	ACC	SE	SP	F1	MCC	AUC
ACPred-FL	8	0.906	0.887	0.924	0.9	0.813	0.969
	16	0.913	0.824	0.902	0.914	0.827	0.971
	32	**0.928**	0.9	**0.954**	**0.926**	**0.857**	**0.982**
	64	0.911	**0.956**	0.866	0.915	0.826	0.977
	128	0.922	0.91	0.934	0.92	0.844	0.978
	256	0.918	0.953	0.883	0.921	0.839	0.979

Table S7. Comparison of the average performance of different filters of CNN on six datasets.

Filter	ACC	SE	SP	F1	MCC	AUC
8	0.913	0.942	0.884	0.916	0.828	0.974
16	0.918	0.915	0.922	0.918	0.837	0.976
32	0.919	0.95	0.888	0.922	0.84	0.978
64	0.923	0.923	0.923	0.923	0.847	0.978
128	0.918	0.939	0.897	0.92	0.838	0.976
256	0.923	0.972	0.874	0.927	0.851	0.984

Supplemental Table 3.2: Experimental Results of different cells

Table S8. Performance comparison of different numbers of hidden cells on ACP530 dataset.

Dataset	Cell	ACC	SE	SP	F1	MCC	AUC
ACP530	8	0.892	0.84	**0.945**	0.883	0.796	0.976
	16	0.902	0.885	0.909	0.902	0.815	0.98
	32	0.923	0.95	0.895	0.923	0.838	0.982
	64	**0.926**	0.94	0.915	**0.925**	**0.851**	**0.983**
	128	0.904	0.91	0.826	0.911	0.817	0.972
	256	0.912	**0.952**	0.883	0.906	0.83	0.978

Table S9. Performance comparison of different numbers of hidden cells on ACP240 dataset.

Dataset	Cell	ACC	SE	SP	F1	MCC	AUC
ACP240	8	0.906	0.915	0.896	0.901	0.862	0.942
	16	0.91	**0.919**	0.902	0.911	0.871	0.944
	32	0.901	0.897	**0.925**	0.906	0.873	0.94
	64	**0.921**	0.912	**0.925**	**0.924**	**0.886**	**0.956**
	128	0.911	0.909	0.914	0.913	0.873	0.945
	256	0.913	0.908	0.919	0.913	0.876	0.946

Table S10. Performance comparison of different numbers of hidden cells on ACP740 dataset.

Dataset	Cell	ACC	SE	SP	F1	MCC	AUC
ACP740	8	0.9	0.911	0.879	0.883	0.85	0.94
	16	0.902	0.895	**0.928**	0.911	0.872	0.945
	32	0.909	**0.923**	0.895	0.909	0.868	0.948
	64	**0.916**	0.918	0.904	**0.913**	**0.883**	**0.953**
	128	0.901	0.94	0.868	0.902	0.855	0.939
	256	0.907	0.92	0.875	0.908	0.866	0.947

Table S11. Performance comparison of different numbers of hidden cells on ACPmain dataset.

Dataset	Cell	ACC	SE	SP	F1	MCC	AUC
ACPmain	8	0.903	0.901	0.906	0.904	0.857	0.941
	16	0.914	0.92	0.908	**0.915**	0.879	0.944
	32	0.902	0.879	0.926	0.902	0.856	0.932
	64	**0.916**	**0.928**	0.889	0.914	**0.88**	**0.945**
	128	0.905	0.871	**0.929**	0.905	0.867	0.94
	256	0.906	0.927	0.873	0.908	0.876	0.941

Table S12. Performance comparison of different numbers of hidden cells on ACPred-Fuse dataset.

Dataset	Cell	ACC	SE	SP	F1	MCC	AUC
ACPred-Fuse	8	0.89	0.852	**0.926**	0.887	0.832	0.93
	16	0.898	0.863	0.913	0.89	0.85	0.936
	32	0.912	0.915	0.909	0.912	0.872	0.942
	64	**0.916**	0.906	0.916	**0.915**	**0.893**	**0.946**
	128	0.915	**0.922**	0.905	**0.915**	0.883	0.945
	256	0.912	0.92	0.906	0.912	0.875	0.943

Table S13. Performance comparison of different numbers of hidden cells on ACPred-FL dataset.

Dataset	Cell	ACC	SE	SP	F1	MCC	AUC
ACPred-FL	8	0.892	0.897	0.921	0.904	0.846	0.938
	16	0.913	0.934	0.89	**0.913**	**0.887**	0.943
	32	0.91	0.9	0.92	0.91	0.871	0.941
	64	0.916	0.89	0.921	0.905	0.872	0.943
	128	**0.918**	0.901	**0.927**	**0.913**	0.878	**0.945**
	256	0.905	**0.92**	0.881	0.906	0.862	0.94

Table S14. Comparison of the average performance of different numbers of hidden cells on six datasets.

Cell	ACC	SE	SP	F1	MCC	AUC
8	0.903	0.901	0.906	0.904	0.857	0.941
16	0.914	0.92	0.908	0.915	0.879	0.944
32	0.902	0.879	0.926	0.902	0.856	0.932
64	0.916	0.928	0.889	0.914	0.88	0.945
128	0.905	0.871	0.929	0.905	0.867	0.94
256	0.906	0.927	0.873	0.908	0.876	0.941

Supplemental Table 3.3: Experimental Results of different critical probabilities

Table S15. Performance comparison of different critical probabilities on ACP530 dataset.

Dataset	Pro	ACC	SE	SP	F1	MCC	AUC
ACP530	0.3	0.891	0.857	0.93	0.917	0.796	0.967
	0.4	0.902	0.927	0.79	0.91	0.828	0.97
	0.5	**0.925**	**0.93**	0.92	**0.926**	**0.854**	**0.973**
	0.6	0.922	0.903	0.918	0.92	0.848	0.969
	0.7	0.916	0.875	**0.96**	0.913	0.841	0.958

Table S16. Performance comparison of different critical probabilities on ACP240 dataset.

Dataset	Pro	ACC	SE	SP	F1	MCC	AUC
ACP240	0.3	0.912	0.925	0.798	0.868	0.771	0.923
	0.4	0.87	0.906	0.853	0.85	0.746	0.889
	0.5	**0.932**	**0.954**	0.896	**0.894**	**0.865**	**0.972**
	0.6	0.928	0.906	0.951	0.89	0.858	0.969
	0.7	0.909	0.841	**0.968**	0.853	0.826	0.943

Table S17. Performance comparison of different critical probabilities on ACP740 dataset.

Dataset	Pro	ACC	SE	SP	F1	MCC	AUC
ACP740	0.3	0.881	0.892	0.87	0.873	0.762	0.91
	0.4	0.88	0.884	0.875	0.878	0.76	0.912
	0.5	0.921	**0.943**	0.907	0.898	0.846	0.967
	0.6	**0.923**	0.879	**0.959**	**0.9**	**0.851**	**0.97**
	0.7	0.912	0.863	0.941	0.881	0.829	0.961

Table S18. Performance comparison of different critical probabilities on ACPmain dataset.

Dataset	Pro	ACC	SE	SP	F1	MCC	AUC
ACPmain	0.3	0.872	0.857	0.91	0.864	0.756	0.892
	0.4	0.902	0.917	0.883	0.884	0.807	0.907
	0.5	**0.915**	0.859	**0.925**	**0.906**	**0.837**	**0.963**
	0.6	0.903	**0.929**	0.859	0.894	0.828	0.96
	0.7	0.902	0.919	0.865	0.883	0.83	0.958

Table S19. Performance comparison of different critical probabilities on ACPred-Fuse dataset.

Dataset	Pro	ACC	SE	SP	F1	MCC	AUC
ACPred-Fuse	0.3	0.894	0.867	0.923	**0.894**	0.816	0.916
	0.4	0.883	**0.936**	0.786	0.876	0.803	0.903
	0.5	**0.911**	0.904	0.909	0.884	**0.834**	**0.958**
	0.6	0.908	0.882	0.92	0.862	0.824	0.956
	0.7	0.909	0.852	**0.941**	0.88	0.826	0.955

Table S20. Performance comparison of different critical probabilities on ACPred-FL dataset.

Dataset	Pro	ACC	SE	SP	F1	MCC	AUC
ACPred-FL	0.3	0.892	0.859	0.912	0.914	0.83	0.911
	0.4	0.92	**0.927**	0.786	**0.931**	0.853	0.946
	0.5	**0.925**	0.913	**0.931**	0.93	**0.859**	**0.951**
	0.6	0.924	0.923	0.918	0.923	0.86	0.95
	0.7	0.916	0.896	0.925	0.913	0.846	0.943

Supplemental Table 3.4: Experimental Results of different component combinations

Table S21. Performance comparison of different component combinations on ACP240 dataset.

Dataset	Component	ACC	AUC	MCC	SE	SP
ACP240	-Ex Transformer layers	0.85	0.73	0.70	0.78	0.75
	-Ex Bi-LSTM modules	0.87	0.84	0.74	0.83	0.81
	-Ex feature fusion mechanisms	0.79	0.74	0.68	0.80	0.76

Table S22. Performance comparison of different component combinations on ACP740 dataset.

Dataset	Component	ACC	AUC	MCC	SE	SP
ACP740	-Ex Transformer layers	0.81	0.76	0.69	0.83	0.75
	-Ex Bi-LSTM modules	0.84	0.86	0.74	0.77	0.85
	-Ex feature fusion mechanisms	0.75	0.73	0.66	0.73	0.79

Table S23. Performance comparison of different component combinations on ACP530 dataset.

Dataset	Component	ACC	AUC	MCC	SE	SP
ACP530	-Ex Transformer layers	0.71	0.76	0.52	0.73	0.70
	-Ex Bi-LSTM modules	0.65	0.70	0.74	0.77	0.85
	-Ex feature fusion mechanisms	0.68	0.77	0.66	0.80	0.58

Table S24. Performance comparison of different component combinations on ACPmain dataset.

Dataset	Component	ACC	AUC	MCC	SE	SP
ACPmain	-Ex Transformer layers	0.73	0.75	0.53	0.76	0.68
	-Ex Bi-LSTM modules	0.66	0.69	0.49	0.67	0.68
	-Ex feature fusion mechanisms	0.64	0.67	0.47	0.71	0.58

Table S25. Performance comparison of different component combinations on ACPred-Fuse dataset.

Dataset	Component	ACC	AUC	MCC	SE	SP
ACPred-Fuse	-Ex Transformer layers	0.72	0.73	0.26	0.65	0.77
	-Ex Bi-LSTM modules	0.68	0.69	0.19	0.68	0.69
	-Ex feature fusion mechanisms	0.62	0.61	0.24	0.58	0.63

Table S26. Performance comparison of different component combinations on ACPred-FL dataset.

Dataset	Component	ACC	AUC	MCC	SE	SP
ACPred-FL	-Ex Transformer layers	0.82	0.87	0.75	0.80	0.82
	-Ex Bi-LSTM modules	0.78	0.84	0.71	0.76	0.74
	-Ex feature fusion mechanisms	0.72	077	0.65	0.70	0.72

Supplemental Table 3.5: Experimental Results of different methods

Table S27. Comparison of eight methods on ACP240 dataset.

Dataset	Model	ACC	AUC	MCC	SE	SP
ACP240	ACP-DA	0.80	0.85	0.65	0.78	0.76
	ACP-DL	0.70	0.82	0.60	0.75	0.72
	ACP- MHCNN	0.85	0.88	0.72	0.82	0.79
	iACP	0.78	0.86	0.68	0.80	0.77
	CL-ACP	0.88	0.90	0.75	0.86	0.85
	ACP-check	0.90	0.92	0.78	0.88	0.87
	TriNet	0.87	0.89	0.75	0.89	0.81
	ACPScanner	0.88	0.91	0.74	0.85	0.85
	ACP-TransLSTM	**0.92**	**0.93**	**0.80**	**0.90**	**0.89**

Table S28. Comparison of eight methods on ACP740 dataset.

Dataset	Model	ACC	AUC	MCC	SE	SP
ACP740	ACP-DA	0.81	0.74	0.58	0.80	0.82
	ACP-DL	0.81	0.89	0.62	0.81	0.80
	ACP- MHCNN	0.86	0.90	0.72	0.89	0.83
	iACP	0.81	0.86	0.61	0.87	0.74
	CL-ACP	0.84	0.91	0.68	0.83	0.85
	ACP-check	0.87	0.92	0.75	0.86	0.88
	TriNet	0.90	0.89	0.82	0.87	0.89
	ACPScanner	0.91	0.96	0.80	0.89	0.90
	ACP-TransLSTM	**0.94**	**0.98**	**0.86**	**0.93**	**0.92**

Table S29. Comparison of eight methods on ACP530 dataset.

Dataset	Model	ACC	AUC	MCC	SE	SP
ACP530	ACP-DA	0.85	0.84	0.73	0.76	0.78
	ACP-DL	0.79	0.78	0.69	0.74	0.80
	ACP- MHCNN	0.73	0.71	0.60	0.69	0.71
	iACP	0.82	0.82	0.75	0.78	0.82
	CL-ACP	0.55	0.39	0.43	0.77	0.39
	ACP-check	**0.93**	**0.96**	**0.85**	0.80	**0.96**
	TriNet	0.91	0.90	0.83	0.85	0.82
	ACPScanner	0.90	0.92	0.84	0.86	0.94
	ACP-TransLSTM	0.91	0.95	0.83	**0.88**	0.95

Table S30. Comparison of eight methods on ACPmain dataset.

Dataset	Model	ACC	AUC	MCC	SE	SP
ACPmain	ACP-DA	0.75	0.75	0.53	0.77	0.73
	ACP-DL	0.53	0.46	0.09	0.86	0.21
	ACP- MHCNN	0.73	0.71	0.46	0.79	0.67
	iACP	0.55	0.47	0.11	0.78	0.32
	CL-ACP	0.45	0.39	0.12	0.67	0.23
	ACP-check	0.78	0.85	0.56	0.80	0.77
	TriNet	0.80	0.79	0.62	0.77	0.79
	ACPScanner	0.79	0.77	0.58	0.76	0.78
	ACP-TransLSTM	**0.87**	**0.89**	**0.67**	**0.89**	**0.80**

Table S31. Comparison of eight methods on ACPred-Fuse dataset.

Dataset	Model	ACC	AUC	MCC	SE	SP
ACPred -Fuse	ACP-DA	0.86	0.85	0.29	0.68	0.89
	ACP-DL	0.82	0.83	0.22	0.68	0.83
	ACP- MHCNN	0.88	0.86	0.29	0.70	0.88
	iACP	0.88	0.76	0.23	0.55	0.89
	CL-ACP	0.85	0.85	0.29	0.70	0.86
	ACP-check	**0.91**	**0.90**	**0.37**	0.73	**0.92**
	TriNet	0.88	0.89	0.36	0.80	0.87
	ACPScanner	0.87	0.88	0.34	**0.81**	0.84
	ACP-TransLSTM	0.88	0.86	0.35	0.79	0.90

Table S32. Comparison of eight methods on ACPred-FL dataset.

Dataset	Model	ACC	AUC	MCC	SE	SP
ACPred -FL	ACP-DA	0.77	0.83	0.56	0.68	0.87
	ACP-DL	0.79	0.84	0.59	0.74	0.84
	ACP- MHCNN	0.77	0.82	0.55	0.67	0.87
	iACP	0.77	0.80	0.49	0.68	0.80
	CL-ACP	0.88	0.94	0.78	0.81	0.94
	ACP-check	0.91	0.94	0.82	0.87	**0.95**
	TriNet	0.88	0.90	0.77	0.86	0.85
	ACPScanner	0.89	0.93	0.76	0.89	0.88
	ACP-TransLSTM	**0.92**	**0.97**	**0.85**	**0.90**	0.92