Localized Regression and Random Forest Modelling Expose Possible Conditional MPST and Copy Number Drivers of Ferroptosis Sensitivity

Ömer Ekmel Kara, MSc.

1.Abstract

Ferroptosis resistance is classically mediated by GPX4 and AIFM2 (FSP1), yet emerging evidence suggests persulfide-generating enzymes such as MPST provide additional protection. Using a pan-cancer dataset, MPST's role in erastin, a potent System Xc^- inhibitor / GPX4 activity suppressor small molecule, resistance was examined. While simple regression linked MPST expression to resistance, this association disappeared with a borderline p value in the multiple linear regression (MLR) models once GPX4 and AIFM2 were included. Stratified models revealed a context-dependent effect: in AIFM2-high cells, MPST emerged as the dominant predictor, surpassing both GPX4 and AIFM2. Random Forest analysis reinforced this finding, and partial dependence plots revealed a synergistic interaction between MPST and AIFM2.

Sliding window regression confirmed amplification of MPST's effect at higher AIFM2 levels. Copy-number (CN) analysis further showed that all significant MPST-containing CN events were completely gains, often co-harboring erastin resistance and few sensitivity corallated genes in high AIFM2 subset, though one unique event implicated MPST alone. These results identify MPST as a conditional effector of ferroptosis resistance, strengthened in AIFM2-high and CN-driven contexts and highlight the segments and co-affected genes which are candidates for effectors of broader interactions affecting the interplay between erastin resistance and ferroptosis.

Contents

2. Key Terms:

MPST; AIFM2 (FSP1); GPX4; Ferroptosis; Erastin; Copy number (CN); CN segment (seg); Z-score; Multiple Linear Regression (MLR); Random Forest (RF); Partial Dependence Plot (PDP); Sliding window regression; Co-amplification; Persulfidase; DepMap

3. Introduction

Conditional sulfur metabolism axis. Ferroptosis resistance is primarily controlled by two independent antioxidant systems: GPX4, which detoxifies lipid peroxides using glutathione, and AIFM2 (FSP1), which regenerates CoQ10 in an NAD(P)H-dependent manner. However, mounting evidence suggests that enzymes producing persulfides and related sulfur species may provide additional layers of protection against lipid peroxidation (Barayeu et al., 2023). Within this sulfur axis, MPST has emerged as a candidate effector whose contribution appears context-dependent, particularly in cells with elevated AIFM2 expression.

In the previous study titled “A Stratified, Lineage-Agnostic Analysis of Persulfide-Related Protein Expression and Erastin Resistance Across Cancer Cell Lines” a significant correlation between MPST protein expression and erastin drug resistance was identified using a simple and multiple linear regression model across a large panel of cancer cell lines (Kara, 2025) . In the current study, LR analysis was repeated with Depmap expression tool, and MLR analysis was repeated with z-score–normalized expression values to improve statistical linearity (Fig. 1A; Fig.2 A, B). However, yet again this association in the linear model was no longer significant in multiple linear regression (MLR) models once GPX4 and AIFM2 protein expression were included as covariates in the full cell set(all) (Fig. 2A, B). Within the high-AIFM2 subset in the MLR model, MPST emerged as the dominant predictor, accounting for the majority of the model’s explanatory variance, with GPX4 contributing modestly and AIFM2 itself showing no significant independent effect (Fig. 2A, B). This highlighted a conditional dependency, where the influence of MPST on erastin resistance becomes most apparent in the context of elevated AIFM2. To further probe potential nonlinear relationships, the MLR analysis model was complemented with a Random Forest (RF). RF modeling consistently identified MPST and GPX4, almost equally, as the most influential predictor of erastin response, particularly in the high-AIFM2 subgroup, where its importance score far exceeded that of AIFM2 (Fig. 2C). Partial dependence plots (PDPs) confirmed this effect and revealed a clear interaction: cells with both high MPST and high AIFM2 exhibited the strongest predicted resistance (Fig. 2D). While results of the RF and PDPs do not establish causality or the direction of dependence, they suggest a synergistic interaction in which MPST’s protective effect is amplified by AIFM2 expression.

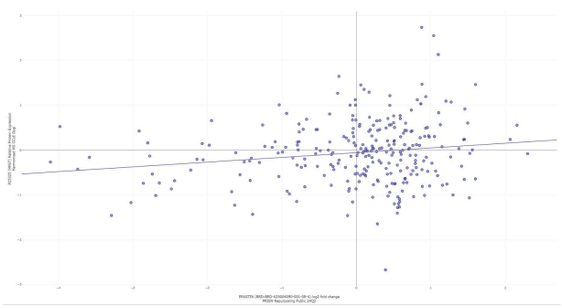
To validate this context-specific trend, a sliding window regression (window size = 40, step = 5) across the ordered dataset (n = 276), with cell lines ranked by AIFM2 protein expression were applied. This analysis revealed a progressive increase in MPST regression slopes at higher AIFM2 levels (Fig.1 B), consistent with the hypothesis that MPST’s protein expression corelation with AIFM2 driven erastin resistance. A similar windowed approach applied to copy number (CN) versus erastin resistance demonstrated stronger confidence intervals, significance levels, and Pearson correlations compared to protein–erastin associations (Fig.1B), indicating a potential genomic basis for these effects.

Finally, all copy number (CN) events (‘segments’, seg) encompassing *MPST* were extracted within the high-AIFM2 subgroup, comprising 37 cell lines in total. The exact number of lines included varied by gene, depending on data availability for downstream analyses. Within this cell subset, 28 CN gain or amplification events, each from a different cell line, were filtered from DepMap CNV data (Supplementary Table 2). Protein-level regressions of co-effected genes in these segments in high AIFM2 cell subset confirmed this trend, identifying 11 positively correlated (including MPST) and only 3 negatively correlated gene (Fig. 3B). Mapping these associations across CN segments revealed multi-gene resistance modules, with each segment containing at least one, other than MPST, but mostly more significantly correlated genes. The sole exception was seg28, where MPST alone accounted for the effect due to the absence of other co-affected genes (Fig. 3B, C). Together, these results indicate that MPST typically operates within CN-driven resistance clusters but can also function independently, with its contribution most pronounced in AIFM2-high contexts.

4.Figures

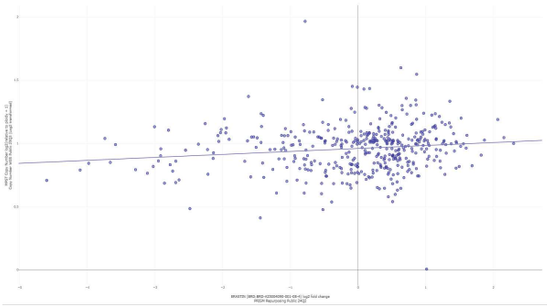
Figure 1

A. LN& WA- MPST Protein Expression vs Erastin Resistance



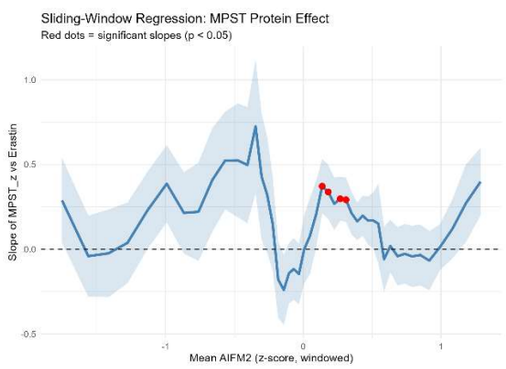
Points	Pearson	Spearman	Slope	Intercept	p-value (linregress)
276	0.162	0.116	1.06e-1	-5.71e-2	6.84e-3

LN & WA- MPST CN vs Erastin Resistance



Points	Pearson	Spearman	Slope	Intercept	p-value (linregress)
432	0.140	0.149	2.35e-2	9.60e-1	3.44e-3

B.



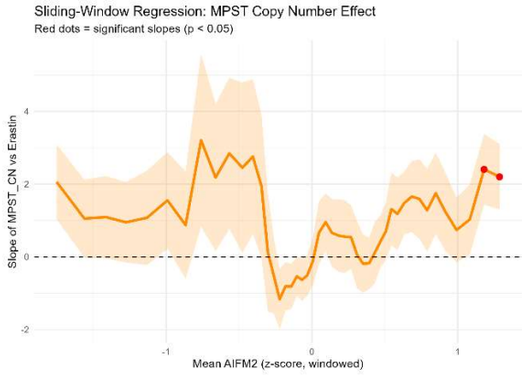
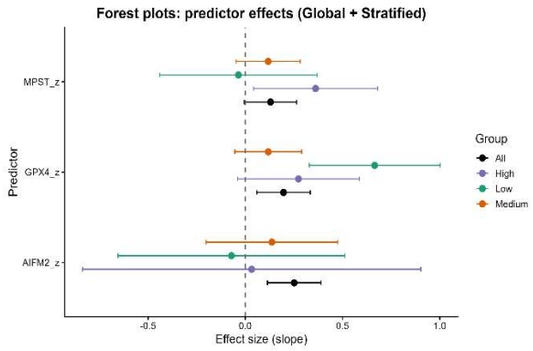


Figure 2. MPST protein and copy number effects on erastin resistance.
A) Protein expression (Gygi z-scores) versus erastin resistance (log2 scores). a1) Simple linear regression of MPST protein expression across all cell lines. a2) Sliding window regression of MPST protein expression versus erastin resistance. Cells were ordered by AIFM2_z, with a window size of 40 and step size of 5. Statistically significant windows ($p < 0.05$) are highlighted as red dots. B) Copy number variation (log2 scores) versus erastin resistance (log2 scores). b1) Simple linear regression of MPST copy number across all cell lines. b2) Sliding window regression of MPST copy number versus erastin resistance, ordered by AIFM2_z (window size = 40, step size = 5). Statistically significant windows ($p < 0.05$) are highlighted as red dots.

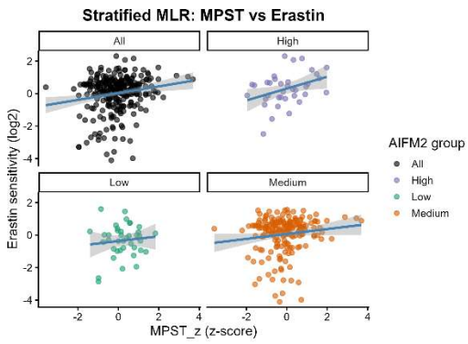
*All sliding window regression results (slope, standard error, t-values, and p-values) for every window are provided in **Supplementary Data 1**.*

Figure 2

A.

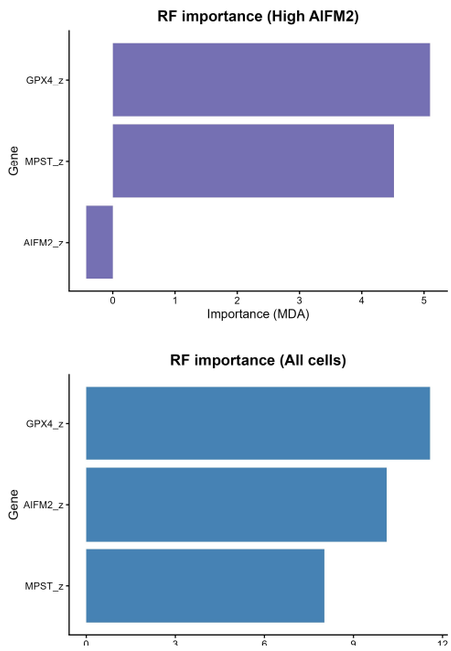


B.

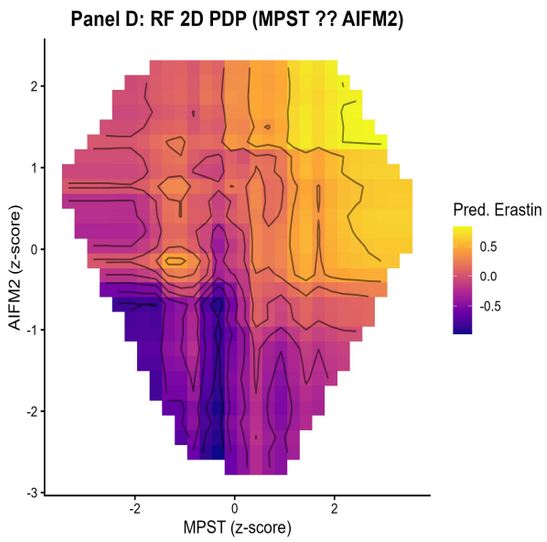


AIFM2_group	SampleSize	estimate	conf.low	conf.high	p.value
All	270	0.130	-0.005	0.264	0.058
High	36	0.362	0.042	0.681	0.028
Low	39	-0.035	-0.440	0.369	0.860
Medium	195	0.117	-0.047	0.282	0.162

C.



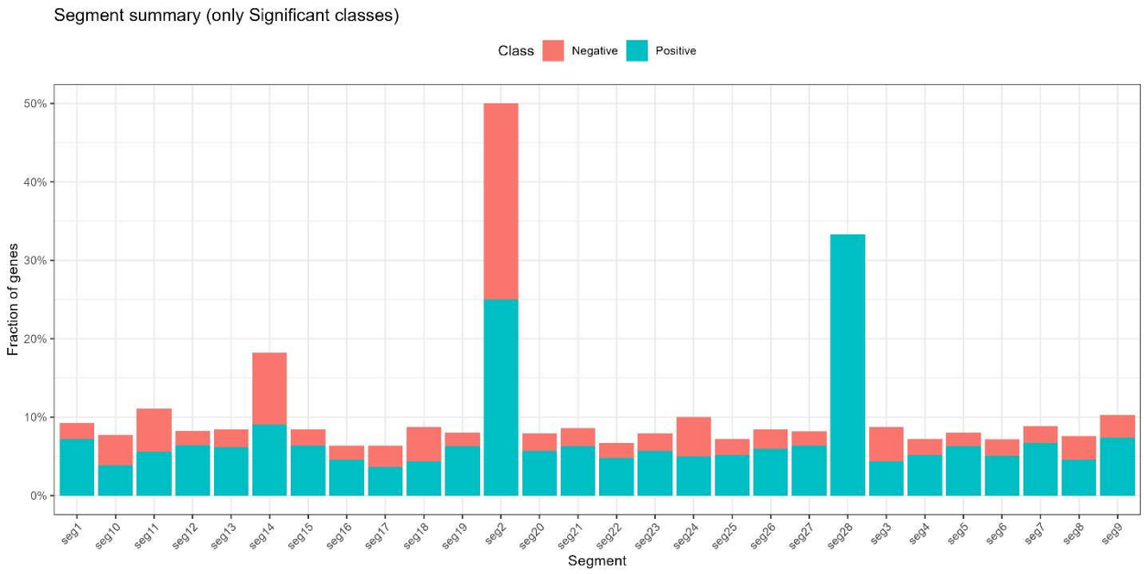
D.



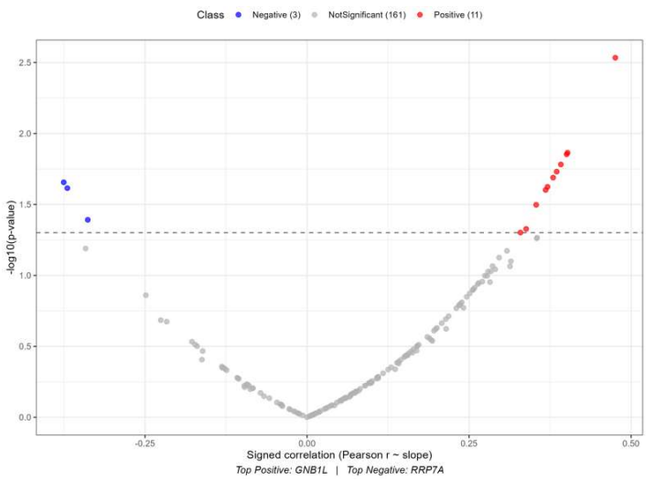
A) Forest plots from the multivariable linear regression (MLR) model. MPST-related coefficients and confidence intervals are highlighted in Fig. 1B. B) Stratified MLR plots showing the relationship between MPST expression (z-score) and Erastin sensitivity (log2) across all cell subsets (*Low*, *Medium*, *High*, and *All*). C) Random Forest variable importance, highlighting the contribution of MPST compared to AIFM2 and GPX4 in all cells and in the high-AIFM2 subgroup. D) Random Forest 2D partial dependence plot (PDP) for the interaction between MPST and AIFM2, shown as a heatmap with contour overlays.

Figure 3

A.

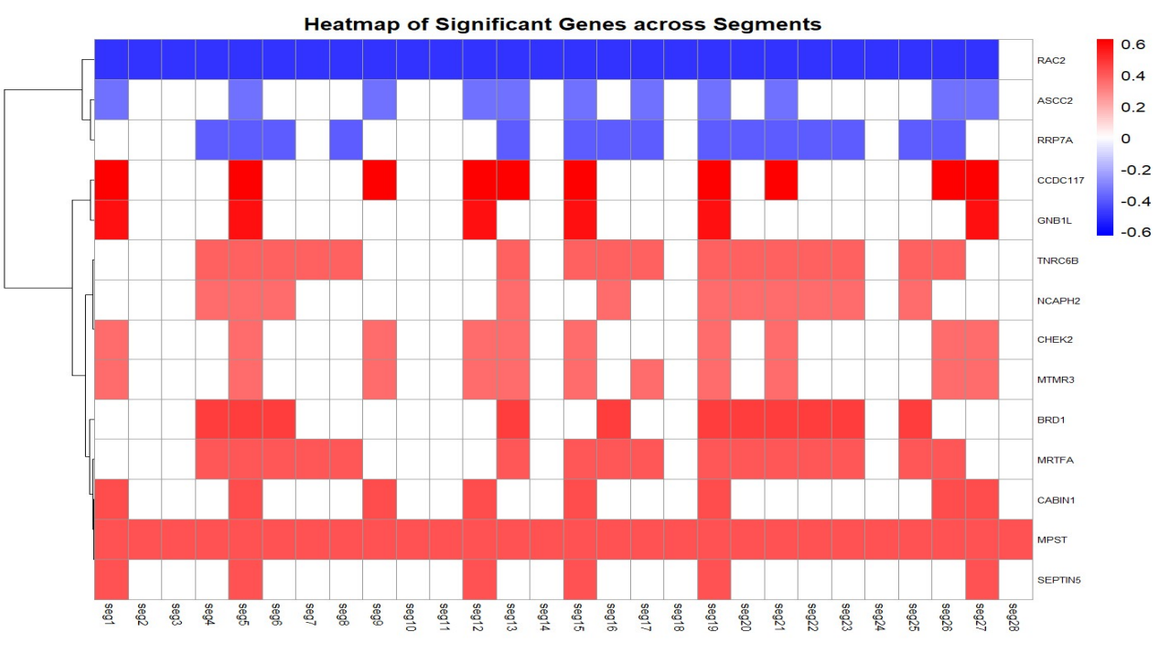


B.



Gene	Class	Slope	PearsonR	Pvalue	R2	N
RRP7A	Negative	-0.3958277	-0.3753196	0.022078239	0.1408648	37
RAC2	Negative	-0.4898627	-0.3697992	0.024276817	0.1367514	37
ASCC2	Negative	-0.3405296	-0.3382655	0.040582264	0.1144235	37
GNB1L	Positive	0.5779646	0.4756427	0.002926857	0.2262360	37
CCDC117	Positive	0.6257046	0.4019575	0.013659257	0.1615698	37
MTMR3	Positive	0.3555092	0.4005524	0.014022841	0.1604422	37
MPST	Positive	0.4248889	0.3916393	0.016524577	0.1533814	37
BRD1	Positive	0.4649131	0.3851629	0.018568852	0.1483504	37
CABIN1	Positive	0.4309200	0.3796926	0.020456511	0.1441664	37
CHEK2	Positive	0.3552945	0.3710482	0.023764112	0.1376768	37
TNRC6B	Positive	0.3762222	0.3680906	0.024992954	0.1354907	37
NCAPH2	Positive	0.3527342	0.3535652	0.031815918	0.1250084	37
SEPTIN5	Positive	0.4194396	0.3379879	0.047053426	0.1142358	35
MRTFA	Positive	0.4050951	0.3292480	0.049899058	0.1084042	36

C.



A)Bar graph of MPST co-affected genes percentage of positive, negative significant correlation with erastin resistance associated with CN events in the AIFM2 high-expression group. Genes are classified as positively or negatively correlated to indicate the slope direction between protein

expression z score and erastin resistance. Table is presented in Supplementary Table 1. B) Volcano plot showing correlations between protein expression and erastin resistance across MPST effected segments in the same cell cohort. C) Heatmap of significant genes across segments; red indicates a significant positive Pearson correlation, while blue indicates a significant negative correlation

5. Methods

5.1 Data Sources and Preprocessing

Erastin sensitivity data.
Drug sensitivity values for erastin were obtained from the PRISM Repurposing dataset (Broad Institute, DepMap 24Q2 release). Sensitivity was expressed as log2 fold-change of cell viability relative to baseline, where more negative values correspond to higher sensitivity. For each comparison, only cell lines with matched proteomic and/or genomic features were retained.

Protein abundance categorization.
Protein abundance values for MPST, AIFM2 (FSP1), and GPX4 were obtained from quantitative mass spectrometry data (CPTAC/DepMap proteomics). Raw intensity values had already been normalized using the Gygi method. For the present analysis, expression values were further standardized by calculating z-scores, and cell lines were categorized into subgroups as follows:

- High expression: $z > 1$ (≈1 SD above the dataset mean, representing overexpression)
- Medium expression: $-1 < z < 1$ (baseline range)
- Low expression: $z < -1$ (underexpression)

Thresholds of ±1 were chosen to exclude borderline cases and to ensure that only biologically meaningful differences in expression were captured, while maintaining adequate statistical power for subgroup comparisons.

Z-score standardization of protein expression values was performed in earlier MLR models. However, due to a coding error, the sample stratification was based on raw Gygi scores rather than standardized values. This introduced unnecessary statistical variance into the previous analysis. Importantly, after correcting the error, the overall trends—particularly the effect of MPST—remained consistent with those reported earlier

Copy number (CN) data.
Gene-level CN values were obtained from DepMap CNV calls, derived from copy number segmentation of whole-exome sequencing data (reported as log2 relative copy number). For locus-specific analyses, segment-level CN events overlapping *MPST* were extracted, together with neighboring genes located within the same CN segments. CN values in DepMap are mean centered across all cell lines, where 0 indicates diploid state, positive values indicate copy number gains, and negative values indicate copy number losses. Both CN–erastin associations and CN–protein associations were assessed to distinguish direct MPST effects from broader co-gained segment effects.

5.2 Linear Models

Multiple Linear Regression (MLR).
Erastin sensitivity (z-scored $\log_2$ fold-change, $n = 276$ cell lines) was modeled against protein expression of MPST, AIFM2 (FSP1), and GPX4. Protein abundance data were derived from quantitative mass spectrometry (DepMap/CPTAC) and normalized using Gygi scores. A global MLR model was first fitted across the entire dataset:

$$\text{Erastin}_{\log 2} \sim \text{MPST}_z + \text{AIFM2}_z + \text{GPX4}_z$$

To evaluate context-specific dependencies, cell lines were stratified by AIFM2 and GPX4 expression levels. States were defined as high ($z > 1$), baseline ($-1 < z < 1$), or low ($z < -1$). Thresholds of ±1 were selected to avoid borderline cases while preserving statistical power and biological interpretability. Subgroup-specific MLR models were then applied, and regression coefficients, 95% confidence intervals, and variance explained (R^2) were estimated.

Windowed linear regression.
To capture continuous variation in predictor influence across AIFM2 contexts, a sliding window regression

was applied. Cell lines were ordered by AIFM2 protein expression, and linear regressions of Erastin sensitivity against MPST expression were computed within moving windows (window size = 40 cell lines, step = 5). Regression slopes and confidence intervals were tracked across windows, allowing detection of progressive changes in MPST's effect with increasing AIFM2 levels. A parallel windowed regression was also applied to MPST copy number (CN) versus Erastin sensitivity, enabling direct comparison of CN–erastin and protein–erastin associations.

5.3 CN analysis

Copy number analysis.
Segment-level copy number (CN) events overlapping MPST in AIFM2 high subset were extracted from the dataset. Neighboring genes within the same CN segments were tested individually at protein expression levels against Erastin sensitivity. Significance thresholds were applied ($p < 0.05$), and correlation directionality was recorded. Genes positively associated with resistance were mapped across CN segments to identify co-gained clusters. Segment-specific effects were then compared: while most segments contained multiple positively correlated genes, only one (seg28) showed a uniquely attributable MPST effect, due to the absence of other significantly associated genes.

5.4 Non-Linear Models

Random Forest modeling and Partial Dependence Plots (PDPs).
To evaluate potential nonlinear effects and predictor interactions, Random Forest (RF) models were constructed using MPST, AIFM2, and GPX4 as input variables, with Erastin sensitivity as the response. Models were fitted with 1000 trees and variable importance scores were extracted (measured as mean decrease in accuracy, MDA). PDPs were generated to visualize predictor-specific effects (MPST, AIFM2, GPX4) and two-way interactions. For the MPST × AIFM2 analysis, 2D PDPs were computed with contour and heatmap overlays, highlighting the predicted effects across the joint predictor space.

5.5 Assistance and Tools

Editing of the manuscript and refinement of figure legends, glossary terms, and structural organization were supported using *ChatGPT* (OpenAI), which was applied exclusively for language editing, summarization, and code annotation. All computational analyses, including data preprocessing, regression modeling, and visualization, were performed in R (version 4.3.3) using standard libraries (e.g., *dplyr, tidyr, ggplot2, randomForest*). Custom R scripts were generated and refined by the author, with occasional code formatting and troubleshooting assistance from *ChatGPT*.

6. Results and Discussion

Across the analyzed cell lines, all 30 MPST-containing copy-number (CN) events corresponded to CN gains or amplifications. In the AIFM2-high subgroup (AIFM2_z > 1), these gains typically encompassed 1–11 proteins positively and 0–3 proteins negatively correlated (z-scores) with erastin resistance. Notably, the JHH cell line (seg28) harbored a CN event including only MPST as a significant correlated gene, yielding a uniquely MPST-dependent phenotype. By contrast, the NCI-H460 cell line (seg1) contained MPST together with RAC2, the sole negatively correlating gene consistently present across amplification events (Fig. 3A; Supplementary Tables 1–2). These models thus provide a valuable framework for dissecting MPST's contribution to erastin resistance both independently (JHH) and in the context of co-affected neighbors (NCI-H460) (Fig. 3D).

Additionally, even though TST protein expression showed only a borderline positive correlation with erastin resistance within the AIFM2 > 1 subgroup in LR and other models, its consistent presence across nearly all MPST-containing CN events is notable (Supplementary figure 1A, B). As TST is located immediately adjacent to MPST on chr22q12.3, its co-amplification may represent a structural feature of these genomic gains (supplementary figure 1A). This raises the possibility that TST could act as a co-dependent persulfidase variable, warranting consideration in future modeling efforts or wet-lab validation studies. Incorporating TST alongside MPST into experimental systems may help clarify whether the observed effects are attributable to MPST alone or to broader co-amplification of neighbouring genes including persulfidase like TST.

These findings align with proteogenomic and pharmacogenomic strategies previously used to link CN alterations, protein dosage effects, and drug response. For example, a proteogenomic study of

ferroptosis regulators integrated CN and protein profiles to infer regulatory mechanisms (Wang et al. 2023), while in AML, combining CN and proteomic data improved drug response prediction (Pino et al.2024). Moreover, proteomics-based drug response modeling has been shown to outperform transcriptomics in many contexts (Zheng et al.2023). Therefore, a further meta-analysis, followed by comparative wet-lab validation, could reveal significant co-dependencies and clarify whether these CN-driven effects represent direct AIFM2 and MPST-specific mechanisms or broader regulatory modules influencing ferroptosis sensitivity.

7. References

Barayeu, U., Schilling, D., Eid, M., Xavier da Silva, T. N., Schlicker, L., Mitreska, N., Zapp, C., Gräter, F., Miller, A. K., Kappl, R., Schulze, A., Friedmann Angeli, J. P., & Dick, T. P. (2023). Hydropersulfides inhibit lipid peroxidation and ferroptosis by scavenging radicals. *Nature Chemical Biology, 19*(1), 28–37. https://doi.org/10.1038/s41589-022-01165-0

Wang, X., Zhang, H., Zhang, M., et al. (2023). Proteogenomic characterization of ferroptosis regulators. *iScience, 26*(5), 106734. https://doi.org/10.1016/j.isci.2023.106734

Pino, J. C., Zheng, Y., Miettinen, T. P., et al. (2024). Mapping the proteogenomic landscape enables prediction of drug response in acute myeloid leukemia. *Nature Cancer, 5,* 205–220. https://doi.org/10.1038/s41590-024-01033-6

Zheng, Y., Zhang, Q., Li, H., et al. (2023). Drug response modeling across cancers: Proteomics vs. transcriptomics. *bioRxiv*. Advance online publication. https://doi.org/10.1101/2023.12.01.123456

8. Supplementary Figures & Tables

Supplementary Table 1. CN Segments Each segment corresponds to an *MPST*-associated CN event identified in the AIFM2-high cell set, shown together with co-affected gene numbers with expression and erastin resistance correlation + or - slope directions.

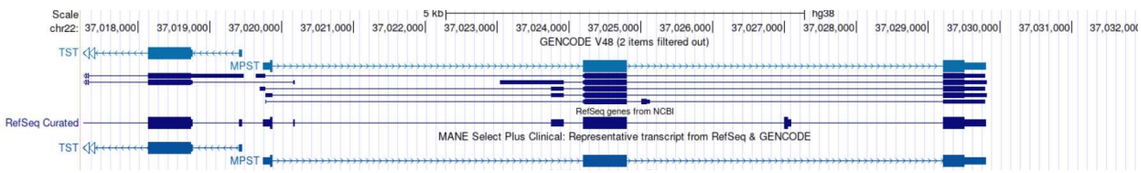
SegmentID	Positive	Negative	NotSignificant	Unknown	TotalGenes	KnownGenes	Pct_Positive	Pct_Negative
seg1	7	2	88	69	166	97	7.22%	2.06%
seg2	1	1	2	8	12	4	25.00%	25.00%
seg3	1	1	21	21	44	23	4.35%	4.35%
seg4	5	2	90	59	156	97	5.15%	2.06%
seg5	11	3	161	113	288	175	6.29%	1.71%
seg6	5	2	91	60	158	98	5.10%	2.04%
seg7	3	1	41	34	79	45	6.67%	2.22%
seg8	3	2	61	42	108	66	4.55%	3.03%
seg9	5	2	61	46	114	68	7.35%	2.94%
seg10	1	1	24	17	43	26	3.85%	3.85%
seg11	1	1	16	15	33	18	5.56%	5.56%
seg12	7	2	100	80	189	109	6.42%	1.83%
seg13	8	3	119	76	206	130	6.15%	2.31%
seg14	1	1	9	13	24	11	9.09%	9.09%
seg15	9	3	130	93	235	142	6.34%	2.11%
seg16	5	2	103	67	177	110	4.55%	1.82%
seg17	4	3	103	66	176	110	3.64%	2.73%
seg18	1	1	21	21	44	23	4.35%	4.35%
seg19	11	3	161	113	288	175	6.29%	1.71%
seg20	5	2	81	54	142	88	5.68%	2.27%
seg21	8	3	117	76	204	128	6.25%	2.34%
seg22	5	2	97	63	167	104	4.81%	1.92%
seg23	5	2	81	54	142	88	5.68%	2.27%
seg24	1	1	18	17	37	20	5.00%	5.00%
seg25	5	2	90	60	157	97	5.15%	2.06%
seg26	7	3	108	81	199	118	5.93%	2.54%
seg27	7	2	101	80	190	110	6.36%	1.82%
seg28	1	0	2	6	9	3	33.33%	0.00%

Supplementary Table 2. the List of the Cell lines and corresponding occurred CN events including MPST in Cell lines subset AIFM2 high.

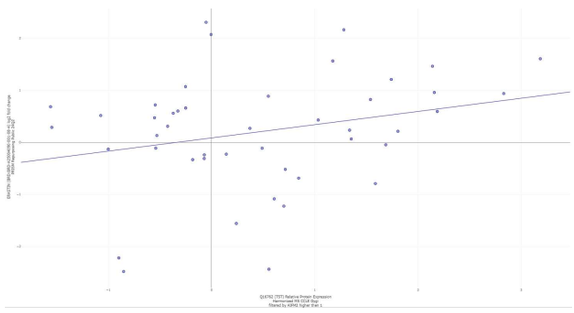
CellLineName	SegmentID	event_type
NCIH460	seg1	Gain
HT115	seg2	Gain
NCIH1792	seg3	Gain
SNU719	seg4	Gain
NCIH226	seg5	Gain
A101D	seg6	Amplification
U2OS	seg7	Amplification
COLO678	seg8	Gain
HCC15	seg9	Gain
IPC298	seg10	Amplification
DU145	seg11	Gain
TCCSUP	seg12	Gain
OV90	seg13	Gain
IGR1	seg14	Amplification
MDAMB231	seg15	Gain
SKCO1	seg16	Gain
LS411N	seg17	Gain
MDAMB436	seg18	Gain
NCIH2052	seg19	Amplification
LS180	seg20	Gain
WM88	seg21	Amplification
C32	seg22	Amplification
TE4	seg23	Amplification
LCLC103H	seg24	Gain
NCIH520	seg25	Gain
SUIT2	seg26	Gain
LS513	seg27	Gain
JHH6	seg28	Amplification

Supplementary Figure 1.

A) GCSC chr22 MPST & TST

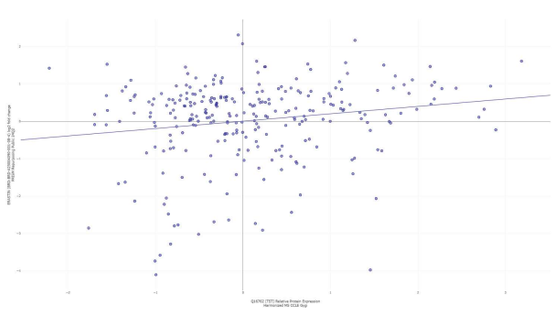


B) LR-TST-AIFM2>1



Points	Pearson	Spearman	Slope	Intercept	p-value (linregress)
45	0.265	0.221	2.54e-1	8.47e-2	7.85e-2

C) LR-TST-All



Points	Pearson	Spearman	Slope	Intercept	p-value (linregress)
276	0.176	0.140	1.97e-1	3.56e-4	3.29e-3

A) GCSC sourced map of TST and MPST on hg38. B) Simple linear regression of TST protein expression vs erastin drug resistance C) Simple linear regression of TST CN score vs erastin drug resistance in all cells

9. Glossary of Key Terms

MPST (Mercaptopyruvate Sulfurtransferase) – An enzyme that produces persulfides, implicated in antioxidant defense and ferroptosis resistance.

AIFM2 (FSP1, Ferroptosis Suppressor Protein 1) – An NAD(P)H-dependent oxidoreductase that regenerates CoQ10 to suppress ferroptosis independently of GPX4.

GPX4 (Glutathione Peroxidase 4) – A key antioxidant enzyme that prevents ferroptosis by reducing lipid peroxides using glutathione.

Ferroptosis – An iron-dependent, non-apoptotic form of regulated cell death caused by lipid peroxidation.

Erastin – A small molecule that induces ferroptosis by inhibiting system Xc^-, limiting cystine uptake and glutathione synthesis.

Copy Number (CN) – The number of copies of a genomic region in a cell; gains, amplifications, loses and deletions can alter gene dosage.

CN Segment (seg) – A contiguous genomic region with uniform copy number state, often encompassing multiple genes.

Z-score – A standardized value indicating how many standard deviations a data point is above or below the mean.

Multiple Linear Regression (MLR) – A statistical model that estimates the relationship between one dependent variable and multiple predictors.

Random Forest (RF) – A machine learning method using decision-tree ensembles to model nonlinear effects and variable importance.

Partial Dependence Plot (PDP) – A visualization showing the marginal effect of one or two predictors on the outcome in a model.

Sliding Window Regression – A method applying regression in overlapping subsets of ordered data to detect context-dependent effects.

Co-amplification – Simultaneous copy number gain of adjacent genes, which may confound attribution of effects to a single gene.

Persulfidase – An enzyme that generates persulfides, providing protective effects against oxidative stress and ferroptosis.

DepMap (Dependency Map Project) – A Broad Institute resource integrating omics and drug-response data across thousands of cancer cell lines.