nature structural & molecular biology

Article

https://doi.org/10.1038/s41594-023-01161-x

The anticancer compound JTE-607 reveals hidden sequence specificity of the mRNA 3′ processing machinery

Received: 11 December 2022

Accepted: 24 October 2023

Published online: 12 December 2023

Check for updates

Liang Liu[1,2,12], Angela M Yu[3,12], Xiuye Wang[1,11], Lindsey V. Soles[1], Xueyi Teng[4], Yiling Chen[4], Yoseop Yoon[1], Kristianna S. K. Sarkan[1], Marielle Cárdenas Valdez[1], Johannes Linder[5], Whitney England[6], Robert Spitale[6,7,8], Zhaoxia Yu[9], Ivan Marazzi[4], Feng Qiao[4], Wei Li[4], Georg Seelig[3,10] ✉ & Yongsheng Shi[1,2] ✉

JTE-607 is an anticancer and anti-inflammatory compound and its active form, compound 2, directly binds to and inhibits CPSF73, the endonuclease for the cleavage step in pre-messenger RNA (pre-mRNA) 3′ processing. Surprisingly, compound 2-mediated inhibition of pre-mRNA cleavage is sequence specific and the drug sensitivity is predominantly determined by sequences flanking the cleavage site (CS). Using massively parallel in vitro assays, we identified key sequence features that determine drug sensitivity. We trained a machine learning model that can predict poly(A) site (PAS) relative sensitivity to compound 2 and provide the molecular basis for understanding the impact of JTE-607 on PAS selection and transcription termination genome wide. We propose that CPSF73 and associated factors bind to the CS region in a sequence-dependent manner and the interaction affinity determines compound 2 sensitivity. These results have not only elucidated the mechanism of action of JTE-607, but also unveiled an evolutionarily conserved sequence specificity of the mRNA 3′ processing machinery.

Almost all eukaryotic mRNA 3′ends are formed through an endonucleolytic cleavage followed by polyadenylation[1,2]. Pre-mRNA 3′ processing is not only an essential step in gene expression, but also an important mechanism for gene regulation. Approximately 70% of human genes produce multiple mRNA isoforms by selecting different PASs, a phenomenon called alternative polyadenylation (APA)[3–5]. Distinct APA isoforms from the same gene can produce functionally different proteins and/or be regulated differently. APA is regulated in a developmental stage- and tissue-specific manner and misregulation of APA contributes to many human diseases[3–5]. It remains poorly understood how APA is regulated in physiological or pathological contexts and pharmacological tools are needed to manipulate APA for research and therapeutic purposes.

PASs are defined by several *cis*-elements, including the AAUAAA hexamer, the U/GU-rich downstream elements and other auxiliary sequences[1,2]. These *cis*-elements are recognized by multiple

[1]Department of Microbiology and Molecular Genetics, School of Medicine, University of California, Irvine, Irvine, CA, USA. [2]Center for Virus Research, University of California, Irvine, Irvine, CA, USA. [3]Department of Electrical and Computer Engineering, University of Washington, Seattle, Seattle, WA, USA. [4]Department of Biological Chemistry, School of Medicine, University of California, Irvine, Irvine, CA, USA. [5]Department of Genetics, Stanford University, Stanford, CA, USA. [6]Department of Pharmaceutical Sciences, University of California Irvine, Irvine, CA, USA. [7]Department of Chemistry, University of California, Irvine, Irvine, CA, USA. [8]Department of Molecular Biology and Biochemistry, University of California, Irvine, Irvine, CA, USA. [9]Department of Statistics, University of California, Irvine, Irvine, CA, USA. [10]Paul G Allen School of Computer Science and Engineering, University of Washington, Seattle, Seattle, WA, USA. [11]Present address: Guangzhou Laboratory, Guangdong, China. [12]These authors contributed equally: Liang Liu, Angela M. Yu. ✉e-mail: gseelig@uw.edu; yongshes@uci.edu

trans-acting factors, including cleavage and polyadenylation specificity factor (CPSF) and cleavage stimulation factor (CstF), which in turn recruit other mRNA 3′ processing factors to assemble the pre-mRNA 3′ processing complex. Pre-mRNA cleavage is carried out by the endonuclease CPSF73 (ref. 6), which, together with CPSF100 and symplekin, forms the nuclease module of the CPSF complex mCF[7]. CPSF73 preferentially cleaves after CA or UA sequences[8]. Although the sequences flanking the CS display distinct and well-conserved nucleotide composition patterns[9–11], it remains unknown what role, if any, these sequences play in pre-mRNA 3′ processing.

CPSF73 has emerged as a drug target for treating a variety of diseases. For example, a number of small molecule drugs for treating *Toxoplasma gondii* (causes toxoplasmosis)[12], African trypanosomes (causes sleeping sickness)[13] and *Plasmodium* spp. (causes malaria)[14], target the CPSF73 homologs in these pathogens. JTE-607 is a small molecule that inhibits the production of multiple cytokines by mammalian cells[15–17] and administration of JTE-607 results in improvements in several inflammatory diseases[15–17]. Furthermore, JTE-607 has anticancer activities and specifically kills the cells of myeloid leukemia and Ewing's sarcoma[18,19]. JTE-607 was also shown to inhibit breast cancer cell migration and invasion[20]. JTE-607 is a prodrug and is hydrolyzed to compound 2 upon entering the cells by the cellular enzyme CES1 (refs. 18,19). Compound 2 specifically binds to CPSF73 near its active site to inhibit its activity[18]. In addition to its potential clinical application, JTE-607 has quickly become an important tool for research[21,22]. However, it is unclear if all pre-mRNA 3′ processing events in the human transcriptome are equally affected by JTE-607 and it is unclear why this compound is active only against specific cancer types.

Although the JTE-607 target, CPSF73, is universally required for pre-mRNA 3′ processing, we have found that JTE-607-mediated inhibition of pre-mRNA 3′ processing is sequence specific both in vitro and in cells. We have identified the CS region as a major determinant of drug sensitivity. Using massively parallel in vitro assays (MPIVAs), we have comprehensively characterized the relationship between the CS sequence and JTE-607 sensitivity, and identified key sequence features that determine drug sensitivity. Using the MPIVA data, we trained a machine learning model, C3PO, that can accurately predict JTE-607 sensitivity of a PAS based on its CS region sequence. We demonstrated that C3PO can help to explain the effect of JTE-607 on PAS selection and transcription termination genome wide. Together, our study not only characterized the properties of an anticancer and anti-inflammation compound, but also revealed a previously unknown sequence specificity of the mRNA 3′ processing machinery.

Sequence-specific mRNA 3′ processing inhibition by JTE-607

To better understand the mechanism of action for compound 2, the active form of JTE-607 (ref. 18), we characterized its effect on pre-mRNA processing in an in vitro cleavage assay using HeLa cell nuclear extract (NE). We first performed in vitro cleavage assays with L3, the PAS of the adenovirus major late transcript, in the presence of dimethyl sulfoxide (DMSO) or increasing concentrations of compound 2 (0.1, 0.5, 2.5, 12.5, 62.5 and 100 μM). Our results showed that the cleavage of L3 PAS was strongly inhibited by compound 2 with a half-maximal inhibitory concentration (IC_{50}) of 0.8 μM (Fig. 1a). To determine whether compound 2-mediated inhibition of pre-mRNA cleavage occurs at the cleavage step and/or the earlier pre-mRNA 3′ processing complex assembly step, we monitored pre-mRNA 3′ processing complex assembly on L3 PAS in the presence of DMSO or increasing concentrations of compound 2 using an electrophoretic mobility shift assay (EMSA). The pre-mRNA 3′ processing complex assembled indistinguishably under all conditions tested (Fig. 1b). These results suggest that compound 2 does not interfere with pre-mRNA 3′ processing complex assembly, but blocks cleavage of the L3 PAS.

We next performed similar in vitro cleavage assays on other PASs. Surprisingly, we found that different PASs displayed different sensitivities to compound 2. For example, cleavage was observed for SVL, the PAS from SV40 late transcript, even at the highest concentration tested of compound 2 with an estimated IC_{50} >100.2 μM (Fig. 1c). Therefore, the IC_{50} of L3 and SVL PASs differ by over 100-fold. Similar to L3, mRNA 3′ processing complex assembly on SVL PAS was not affected by compound 2 (Extended Data Fig. 1). In total we performed in vitro cleavage assays with 40 different PASs and found that their IC_{50} values varied widely (Fig. 1d). To begin to understand the molecular basis for such variations, we first asked whether the compound 2 sensitivity of a PAS is determined by its strength, that is, the efficiency by which it is processed by the pre-mRNA 3′ processing machinery. We measured the percentage of pre-mRNA cleaved in vitro in the absence of compound 2 and compared this value, designated as PAS activity, with its IC_{50}. Our results detected poor correlation between the two measurements ($r = 0.30$) (Fig. 1d and Supplementary Table 1). We concluded that the cleavage of different PASs displays differential sensitivities to compound 2 in vitro and that the drug sensitivity of a PAS is not determined by its strength.

CS sequence is a major determinant of JTE-607 sensitivity

As PASs display sequence-dependent sensitivity to compound 2 in vitro, we next wanted to map the specific PAS region(s) that determine its drug sensitivity. To this end, we divided the PAS sequence into three regions: the AAUAAA hexamer and upstream sequence, the CS region (20 nucleotide (nt) regions centered at the cleavage site) and the downstream sequence (Fig. 2a). Among PAS sequences we tested previously, L3 (IC_{50} = 0.8 μM; Fig. 1a) and SVL (IC_{50} = 100.2 μM; Fig. 1c) showed the lowest and the highest resistance to compound 2, respectively. Therefore, we constructed a series of chimeric PASs between them by replacing one or more of the three regions in one PAS by their counterparts in another, and measured their IC_{50} values as described above. Replacing the upstream sequence of L3 PAS with that of SVL did not result in a major change in IC_{50} (chimera 1, IC_{50} = 2.1 μM; Fig. 2a and Extended Data Fig. 2a). However, replacing both the upstream sequence and the CS of L3 with those of SVL dramatically increased the resistance to compound 2 (chimera 2, IC_{50} = 89.6 μM; Fig. 2a and Extended Data Fig. 2b), suggesting that the CS region plays a major role. On the other hand, replacing the upstream sequence of SVL with that of L3 led to a significant decrease in drug resistance (chimera 3; Fig. 2a and Extended Data Fig. 2c), although its IC_{50} (39.5 μM) was still almost 50-fold higher than that of L3. Replacing both the upstream sequence and CS of SVL with those of L3 led to an almost 15-fold decrease in IC_{50} (chimera 4, IC_{50} = 6.7 μM; Fig. 2a and Extended Data Fig. 2d), again highlighting a major role for the CS region. By contrast, the downstream sequence did not seem to play a major role (compare L3 and chimera 4 or SVL and chimera 2; Fig. 2a). Given the major impact of the CS region on compound 2 sensitivity in both backgrounds, we swapped the CS regions alone between L3 and SVL. The results showed that replacing the L3-CS region with that of SVL changed its IC_{50} to 47.8 μM, an almost 60-fold increase (L3-SVL CS, Fig. 2a.b). Even more dramatically, the opposite change in SVL reduced its IC_{50} to 0.8 μM (SVL-L3 CS; Fig. 2a,c), identical to that of L3. These results demonstrated that the CS region is a major determinant of compound 2 sensitivity in both backgrounds. In addition, the upstream sequence also contributes to the drug sensitivity in a context-dependent manner, whereas the downstream sequence does not appear to play a significant role. Therefore, we focused on the CS region for the rest of the present study.

Characterization of the sequence–drug sensitivity relationship by MPIVA

We next wanted to comprehensively define the relationship between the CS region sequence and compound 2 sensitivity. To this end, we designed

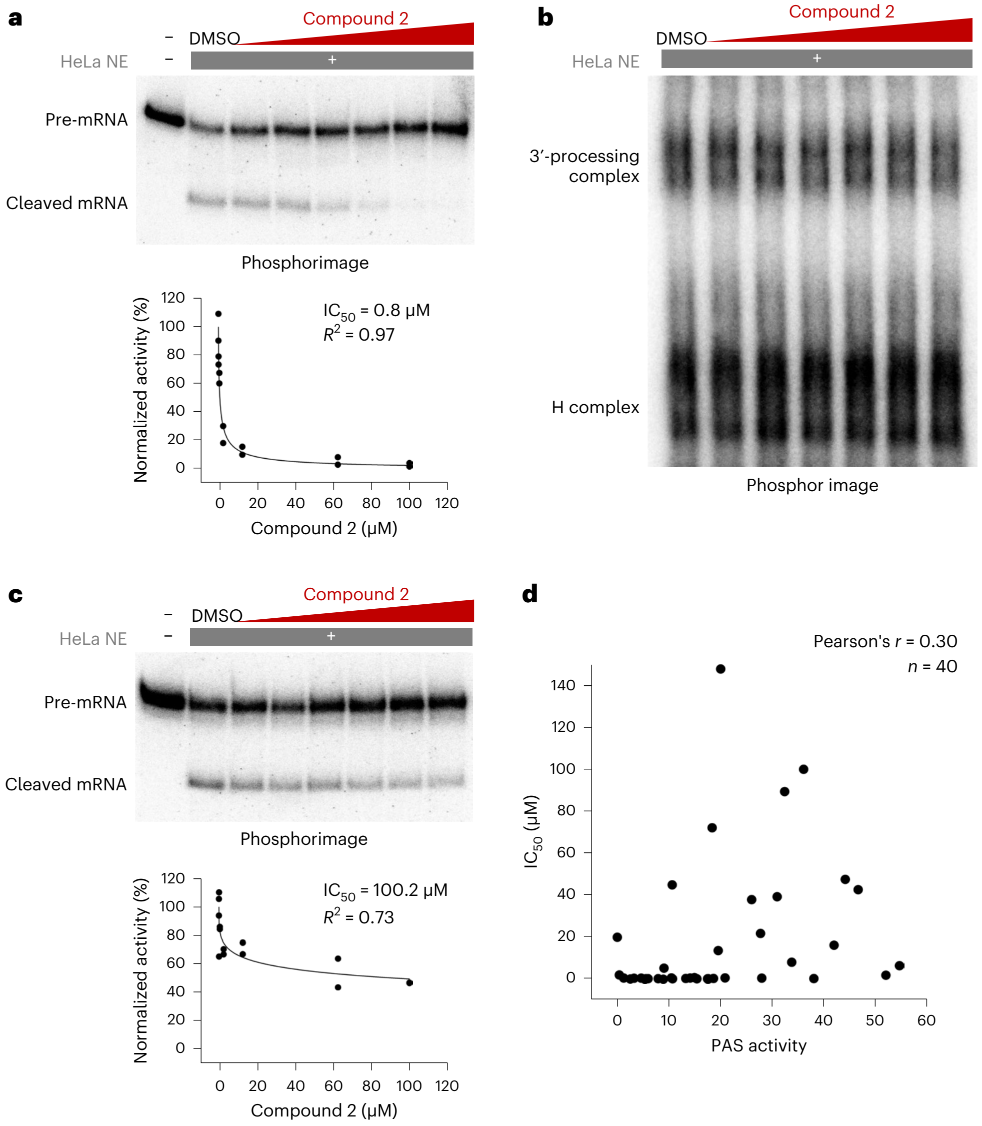


Fig. 1 | Compound 2-mediated inhibition of mRNA 3′ processing in vitro is sequence dependent. a, In vitro cleavage assay on L3 PAS with increasing concentrations of compound 2 and its IC_{50} quantification. Radiolabeled RNAs from the reactions were extracted and resolved on 8 M urea gel and visualized by phosphor imaging. The compound 2 concentrations used are 0.1, 0.5, 2.5, 12.5, 62.5 and 100 μM (*n* = 2 biological replicates and both measurements are shown as dots). **b**, EMSA with L3 PAS in the presence of increasing concentrations of compound 2. The concentrations used are the same as those used in **a**. **c**, In vitro cleavage assay on SVL PAS with increasing concentration of compound 2 and its IC_{50} quantification (similar to **a**). **d**, PAS activity and IC_{50} correlation of 40 in vitro tested PASs.

an MPIVA strategy (Fig. 3a). Using L3 (sensitive) or SVL (resistant) PAS as backbones, we replaced the original cleavage site sequence with a YA (Y is U or C), which is the preferred cleavage site sequence for CPSF73, and randomized the 23-nt flanking sequence. Exchanging UA or CA at the cleavage sites in L3 or SVL did not have a significant effect on compound 2 sensitivities (Extended Data Fig. 3a,b). These two libraries, called L3-N23 and SVL-N23, contained ~3 million PAS variants each and were transcribed into RNAs. The RNA pools were used for in vitro cleavage and polyadenylation assays in the presence of DMSO (control) or compound 2, including low (0.5 μM), medium (2.5 μM) and high (12.5 μM) concentrations. As shown in Fig. 3b, the PAS RNA pool was efficiently cleaved in vitro in the presence of DMSO and the cleavage efficiency gradually decreased in the presence of increasing concentrations of compound 2. The starting PAS RNA pool and the cleaved RNA pools under different conditions were subjected to high-throughput sequencing using the Illumina platform (Fig. 2b). The 12-nt sequence upstream of the CSs in the sequencing reads was used to identify the corresponding full-length sequence and only unambiguously identifiable sequences were kept for further analyses (Fig. 3a; see Methods for details). For each variant, a resistance score was calculated as the log(ratio) between its frequency in compound 2-treated samples and that in DMSO-treated samples. As shown in Fig. 3c and Extended Data Fig. 3c, the resistance scores of all variants were concentrated in a narrow peak centered at ~0 at low compound 2 concentrations (L3: −0.04 ± 0.31; SVL: −0.05 ± 0.28) but diverged more at high inhibitor concentrations (L3: −0.12 ± 0.53; SVL: −0.09 ± 0.43), suggesting that, as expected, drug sensitivities are better distinguished at higher drug concentrations. Furthermore, we compared the resistance scores of all variants and their cleavage efficiency (log(ratio) between the frequency of a PAS variant in library 2 and that in library 1) and found that there was no significant correlation (Fig. 3d and Extended Data Fig. 3d), which was consistent with Fig. 1d. Thus, both our low-throughput in vitro assays and high-throughput screen results demonstrated that the compound 2 sensitivity of a PAS is not dependent on its strength.

Based on the resistance scores in the high compound 2 concentration condition, we obtained a list of the top 1,000 most sensitive

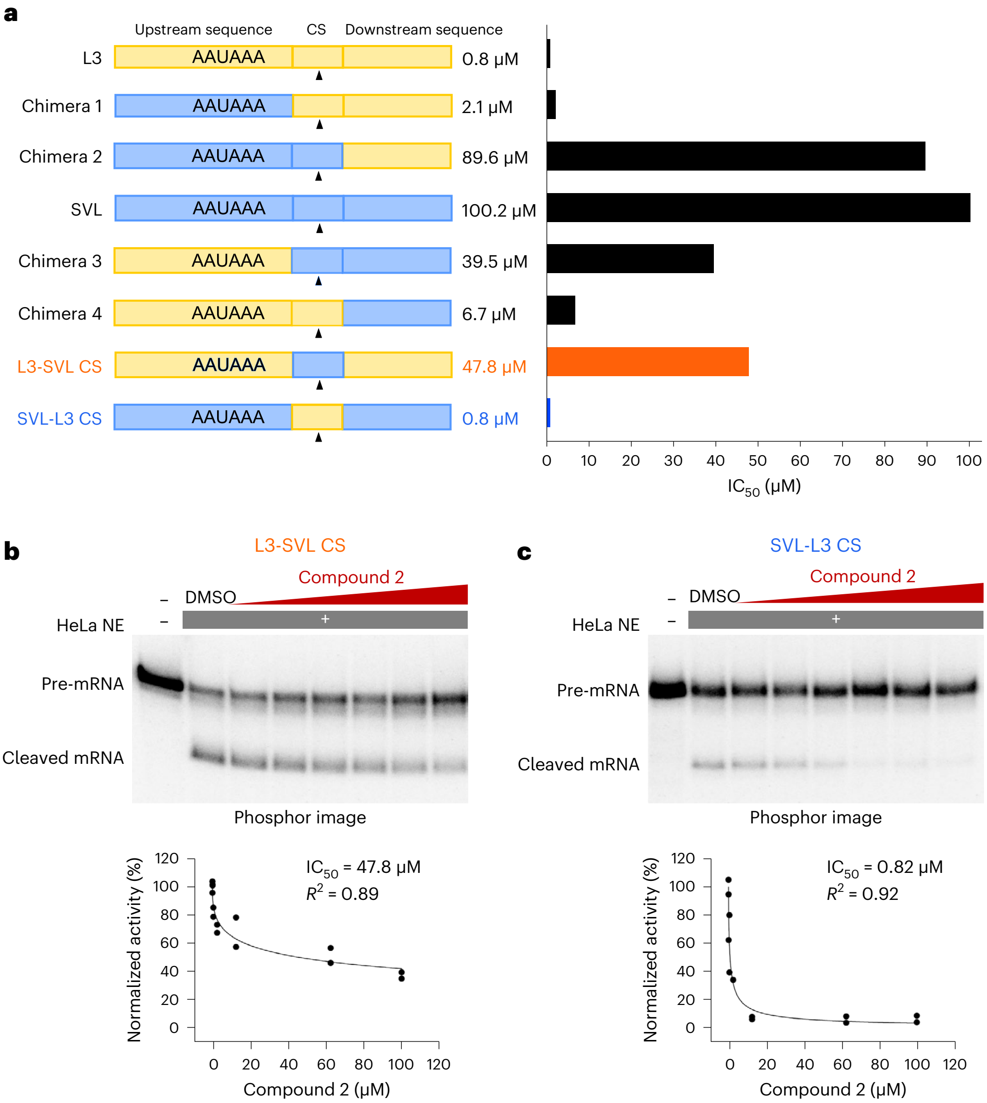


Fig. 2 | CS region is a major determinant of compound 2 sensitivity. **a**, A diagram of L3, SVL and their chimeras. Their corresponding IC_{50} values were plotted on the right. The black triangles denote the cleavage position YA (Y is U or C). **b**, In vitro cleavage of L3-SVL CS with increasing concentration of compound 2, similar to Fig. 1a,c. **c**, In vitro cleavage of SVL-L3 CS with increasing concentrations of compound 2 ($n = 2$ biological replicates and both measurements are shown as dots).

and resistant PASs from both the L3-N23 and the SVL-N23 libraries. We selected four variants in each background, two sensitive and two resistant, and tested them using in vitro cleavage assay and our data validated the screen results (Fig. 3e and Extended Data Fig. 3e,f). It was noted that some of the variants (for example, Fig. 3e, top left) were more sensitive to compound 2 than the original L3 whereas other variants displayed greater resistance than SVL (for example, Fig. 3e, bottom right), indicating that our screens selected variants with a wide range of drug sensitivities. It is interesting that the CS region of sensitive and resistant PASs showed distinct patterns. The CS regions of sensitive L3 variants are generally G/U rich, especially in the region upstream of the cleavage site (Fig. 3f, top). In contrast, resistant CSs contained alternating U-rich and A-rich sequences mainly in the region upstream of the cleavage site (Fig. 3f, bottom). Very similar patterns were observed on an SVL background (Fig. 3g), suggesting that the CS region sequence can determine compound 2 sensitivity independent of other regions. Consistent with the nucleotide compositions, our motif analyses of the sensitive and resistant variants detected U/G-rich and A/U-rich motifs, respectively, in both the L3 and the SVL libraries (Extended Data Fig. 4a,b). These results defined the key sequence features in the CS region that determine compound 2 sensitivity.

Predict JTE-607 relative sensitivity by machine learning

We next used our MPIVA data to train a machine learning model with the goal of predicting compound 2 sensitivity of any PAS based on its CS region sequence. Our model, called cleavage and counteraction with compound 2 on polyadenylation outcomes (C3PO), is a three-layer convolutional neural network (CNN) that is based on the Optimus 5′-architecture that we have previously used to predict polysome profiles from 5′-untranslated region (UTR) sequences (Fig. 4a and Methods)[23]. C3PO uses the 25-nt CS sequences as inputs and predicts compound 2 sensitivity, which is calculated as the log(ratio) between each variant's percentage representation in the DMSO-treated and compound 2-treated libraries (Fig. 3a). C3PO was trained on the processed MPIVA datasets from both the L3 and the SVL RNA contexts and model performance was assessed on held-out variants from both RNA contexts. We used the variants with high-read coverage in the input

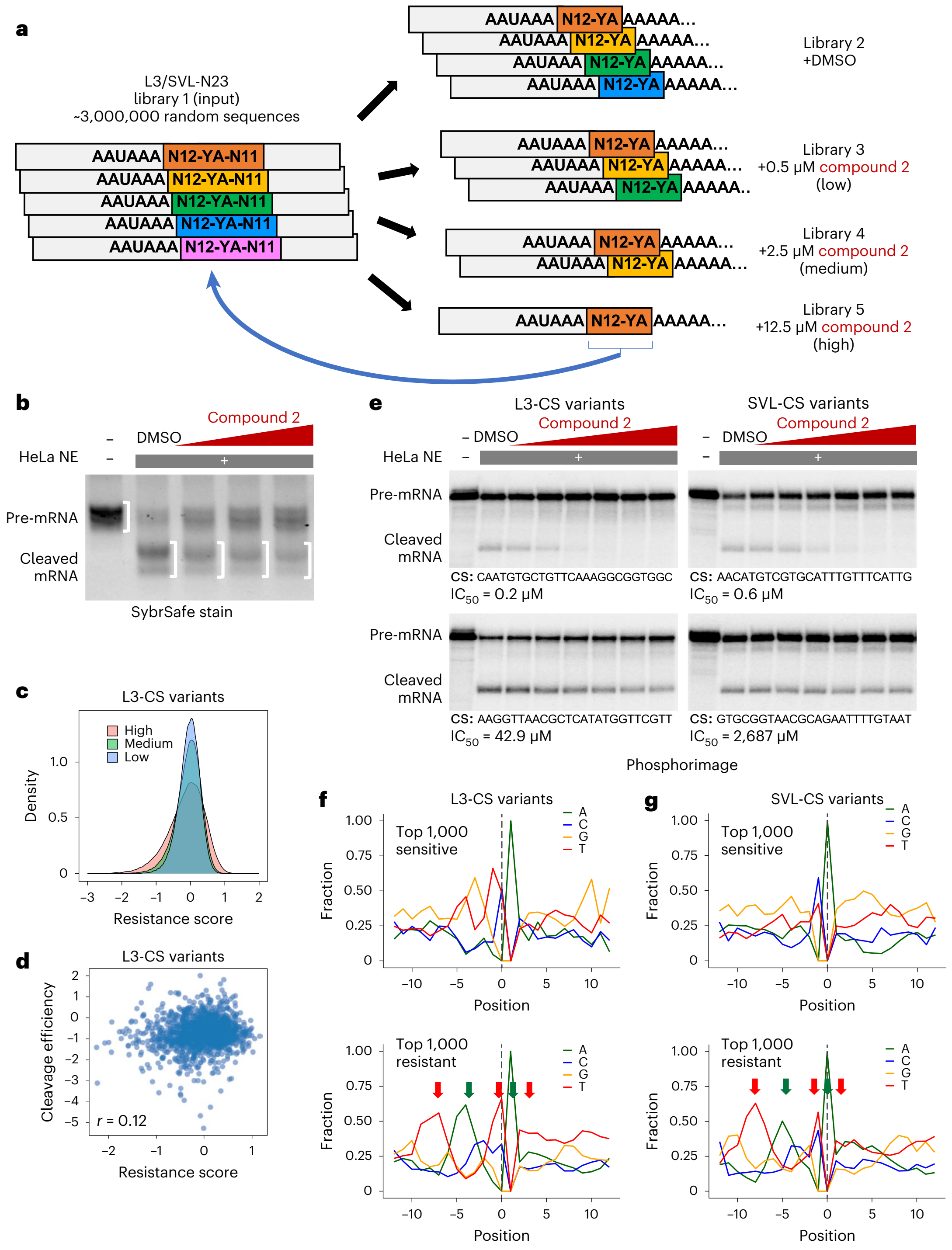


Fig. 3 | Determine sequence specificity for compound 2 sensitivity by MPIVA. **a**, Design of the randomized CS sequence libraries and the MPIVA assay. Each box represents a sequence variant. N, random nucleotide; YA, cleavage position (Y is U or C). The N12 sequence in the sequencing reads was used to identify the corresponding full-length sequence in the library (as shown by the arrow). **b**, The randomized sequence library L3/SVL-N23 transcribed into RNAs and used for in vitro cleavage/polyadenylation assays in the presence of 0.5, 2.5 and 12.5 μM compound 2. The RNAs from these reactions were amplified by RT–PCR and resolved on an agarose gel. The RNA species were marked on the left. The white half-brackets mark the regions on the gel that were extracted and amplified for sequencing ($n = 2$ biological replicates). **c**, A density plot for the resistance scores of all variants in the L3-N23 library. The low, medium and high groups represent the screens in the presence of 0.5, 2.5 and 12.5 μM compound 2 as shown in **b**. **d**, A scatter plot comparing the cleavage efficiency log(frequency in library 2/frequency in library 1) and the resistance score (log(frequency in library 5/frequency in library 2)) of L3-CS variants. The Pearson's correlation is shown. **e**, Examples of validation experiments using in vitro cleavage assays for variants from both L3- and SVL-N23 libraries ($n = 2$ biological replicates). **f**, Nucleoside distribution of L3-CS variants for the top 1,000 most sensitive and resistant sequences. **g**, Nucleoside distribution of SVL-CS variants for the top 1,000 most sensitive and resistant sequences. T- and A-rich regions were marked with red and green arrows, respectively.

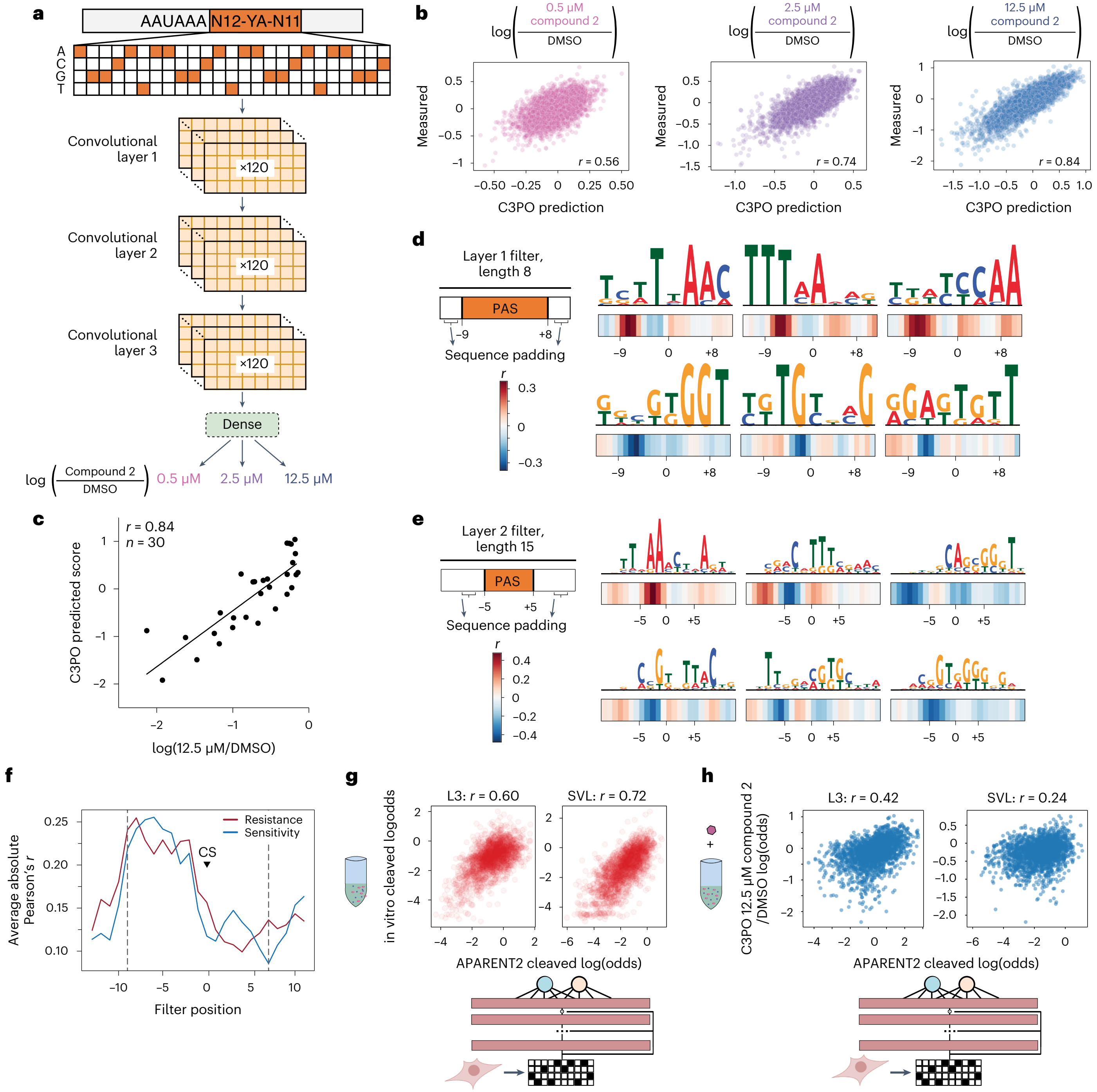


Fig. 4 | C3PO architecture, performance and layer feature analyses. **a**, The model taking 25-nt RNA sequences immediately downstream of the core hexamer and predicting three doses of compound 2 drug sensitivity by predicting the log(ratio) of percentage reads in a drug-treated sample to a DMSO-treated sample. **b**, Scatter plots of C3PO performance on predicting drug sensitivity at three compound 2 doses on test sequences. Test sequences included equal number of sequences derived from both the L3 and the SVL RNA contexts. **c**, A scatter plot comparing the resistance scores predicted by C3PO and those measured experimentally. **d**,**e**, Convolutional layer 1 (**d**) and layer 2 (**e**) max filter activations with the highest Pearson's correlation with 12.5 µM compound 2 predictions. Sequence logos are plotted on top of the per-position absolute value of Pearson's correlations with 12.5 µM compound 2 sensitivity predictions. Additional layer 1 and 2 filters are reported in Extended Data Fig. 4c,d. **f**, Plot of average of all layer 1 filters' absolute value of Pearson's correlation with 12.5 µM compound 2 predictions across all positions. These are split into Pearson's correlation values associated with resistance or sensitivity. The dashed gray lines indicate positions at the edge of the sequence padding. The position of the CS is marked and note that preceding filters may overlap with the designed canonical cut sites. **g**, Scatter plots of RNA cleavage log(odds) measured in vitro calculated from input and DMSO libraries versus those from APARENT2 predictions. **h**, Scatter plots of compound 2 resistance predicted by C3PO and the cleavage efficiency predicted by APARENT2.

and DMSO-treated data (libraries 1 and 2) as our test set to minimize the impact of measurement noise (Methods). C3PO performed better on higher doses of compound 2 with Pearson's *r* of 0.56, 0.74 and 0.84 for 0.5 µM, 2.5 µM and 12.5 µM, respectively. We explored variations of convolution-based machine learning architectures (Supplementary Table 2) and this trend was consistent. This was expected because drug resistance is better detected at higher drug doses (Fig. 3c). Owing to the better model performance at the highest dose of 12.5 μM, we focused further analyses on this regimen.

To test the performance of C3PO, we compared the compound 2 resistance scores (log(12.5 μM/DMSO)) of 30 distinct PASs measured by in vitro cleavage assays as shown in Fig. 1d (PASs that contain the same CS region sequences were combined to avoid redundancy) and those predicted by C3PO. The C3PO predictions showed strong and positive correlation with experimental measurements with a Pearson's *r* of 0.84 (Fig. 4c and Supplementary Table 3). This is very similar to its performance on the MPIVA dataset (compare Fig. 4c with Fig. 4b, 12.5 μM panel). These results strongly suggest that C3PO can accurately predict compound 2 sensitivity of PAS sequences in vitro.

We next wanted to identify sequence motifs that predict compound 2 sensitivity by extracting filter position weight matrices (Fig. 4a). The position-specific effect on compound 2 sensitivity of each filter was quantified by measuring the correlation with drug sensitivity at each position across the CS region. Filters associated with higher resistance (dark-red color) learned motifs that were A/U rich, whereas lower resistance filters (dark blue) typically learned motifs with higher G/U content (Fig. 4d). Sequence motifs strongly associated with compound 2 sensitivity predictions are positioned such that they begin upstream of the CS (Fig. 4d–f and Extended Data Fig. 4c). Layer 2 filters learn to use combinations of layer 1 filters for predictions of drug sensitivity. The 15-mers learned by layer 2 filters also showed A/U-rich and G/U-rich motifs for resistant and sensitive PASs, respectively (Fig. 4e and Extended Data Fig. 4d). It is interesting that both resistance- and sensitivity-associated motifs are enriched in the region upstream of the CS (Fig. 4f).

Given the known function of RNA secondary structures in pre-mRNA 3′ processing[24], we investigated its potential impact on compound 2 sensitivity. We compared the minimum free energy (MFE) structures for the top 10,000 resistant and sensitive sequences (Extended Data Fig. 5a,b). The differences between Δ*G* (Gibbs free energy) values for the resistant and sensitive sequences were modest, but statistically significant with *P* values of $<2.2 \times 10^{-308}$ and 1.58×10^{-26} for L3 and SVL, respectively. The difference between base-pairing probabilities for resistant and sensitive sequences also show different global patterns between the L3 and SVL backbones, indicating that background-specific secondary structural features may contribute to drug sensitivity (Extended Data Fig. 5c,d). Taken together with C3PO's ability to accurately predict compound 2 sensitivity with sequence alone, our results suggest that sequence is the primary determinant of compound 2 sensitivity whereas secondary structure may play a minor role.

We further explored the usage of machine learning models to characterize compound 2 sensitivity and its relationship with processing efficiency. First, we compared the cleavage efficiency measured by our MPIVA assays with that predicted by APARENT2 (ref. 25), a highly accurate deep learning model for predicting cleavage/polyadenylation efficiency that was trained using massively parallel reporter assays in mammalian cells. We saw good correlation between APARENT2-predicted cleavage efficiency and our MPIVA data with a Pearson's *r* of 0.60 for the L3 background and 0.72 for the SVL background (Fig. 4g). These results suggest that the CS region sequence can have a significant impact on cleavage efficiency and the cleavage efficiency values measured by our MPIVA system are highly consistent with measurements obtained in cells. Finally, we compared the resistance score predicted by C3PO with the cleavage efficiency predicted by APARENT2 for all CS variants and observed poor correlation with Pearson's $r = 0.42$ and 0.24 for L3 and SVL, respectively (Fig. 4h). This is consistent with our in vitro cleavage assay (Fig. 1d) and MPIVA results (Fig. 3d and Extended Data Fig. 3d) and provided further evidence that the compound 2 sensitivity of a PAS is not dependent on its strength.

We have also trained C3PO on L3-N23 and SVL-N23 libraries separately. These models achieved similar prediction performance on variants in their respective backgrounds (Supplementary Fig. 6a,d) and identified similar motifs for sensitive and resistant sequences (Extended Data Fig. 6c,f). However, their performance decreased when predicting compound 2 sensitivities in another background (Extended Data Fig. 6a,d). In addition, although the sequence upstream of the CS plays the most important role for the L3-N23-based model (Extended Data Fig. 6b) and the combined model (Fig. 4f), sequences immediately downstream of the CSs are also important for the SVL-N23-based model (Extended Data Fig. 6e). There were also noticeable differences between PASs with CA or UA CS sequence (Extended Data Fig. 7c–h). Finally, we note that high concentrations of compound 2 treatment activated cleavage downstream of the normal CS, albeit at extremely low frequencies (Extended Data Fig. 7a,b), so should have little impact on the modeling of overall drug sensitivity.

Sequence-specific effect of JTE-607 in human cells

To determine whether the sequence-specific sensitivity to compound 2 observed in vitro was true in cells, we performed two genome-wide analyses. First, we analyzed the global APA profiles of human HepG2 cells treated with DMSO (control) or JTE-607 for 4 h using PAS sequencing (PAS-seq), a high-throughput RNA 3′sequencing method for quantitatively mapping RNA polyadenylation[26]. JTE-607 treatment induced significant APA changes in 921 genes, of which 847 genes (92%) shifted from a proximal PAS to a distal one (blue dots; Fig. 5a and see Methods for details). An example was shown in Fig. 5b: the proximal PAS was predominantly used for *Ptp4a1* transcripts in DMSO-treated cells. However, polyadenylation shifted to a distal PAS in JTE-607-treated cells, leading to 3′-UTR lengthening; 74 genes showed APA changes in the opposite direction (red dots; Fig. 5a), as exemplified by *Paqr8* (Fig. 5c).

Why did JTE-607 induce the opposite APA changes in different groups of genes? Given our finding that JTE-607-mediated inhibition of mRNA 3′ processing is sequence specific, we hypothesized that JTE-607 treatment would decrease the relative usage of the more sensitive PASs in a given gene whereas that of resistant PASs would be less impacted, leading to a net shift to more resistant PASs. To test this hypothesis, we predicted the resistance scores of all annotated PASs in the human genome using C3PO and compared the scores of the proximal and distal PASs of the 921 genes that displayed significant APA shifts in JTE-607-treated cells. It is interesting that, for the 847 genes that showed a shift to the distal PAS in JTE-607-treated cells, their proximal PASs are significantly more sensitive to JTE-607 than their distal ones ($P < 2.2 \times 10^{-16}$, Student's *t*-test, Cohen's $d = -0.38$; Fig. 5d, left). The opposite trend was observed for the 74 genes that showed a distal-to-proximal shift ($P = 0.0046$, Student's *t*-test, Cohen's $d = 0.47$; Fig. 5d, right). Therefore JTE-607 indeed inhibited the relative usage of more sensitive PASs, resulting in higher relative usage of resistant PASs. These data confirmed that JTE-607 modulates PAS selection globally in a sequence-dependent manner in human cells and that JTE-607-induced APA changes are influenced by the relative drug sensitivities of the alternative PASs.

In addition, we also monitored transcription termination by nascent RNA sequencing using 4-thiouridine-labeled RNA (4sU-seq) in HepG2 cells treated with DMSO or JTE-607. As mRNA 3′ processing is coupled to transcription termination, transcription termination efficiency at PAS can be used as a proxy for mRNA 3′ processing efficiency[27]. Our 4sU-seq analyses showed that JTE-607 treatment induced a global transcription termination defect (Fig. 6a). However, the levels of JTE-607-induced transcription read through (RT) varied widely at different PASs (Fig. 6a). For example, RT increased dramatically downstream of the PAS of the *Eif4ebp1* gene (Fig. 6b, left) whereas little change was observed for the *Cox8A* gene (Fig. 6b, right). Thus, 4sU-seq data further demonstrated that mRNA 3′ processing displayed sequence-specific sensitivity to JTE-607 in human cells.

We then tested whether JTE-607-induced, gene-specific transcription termination defects can be attributed to the differential drug

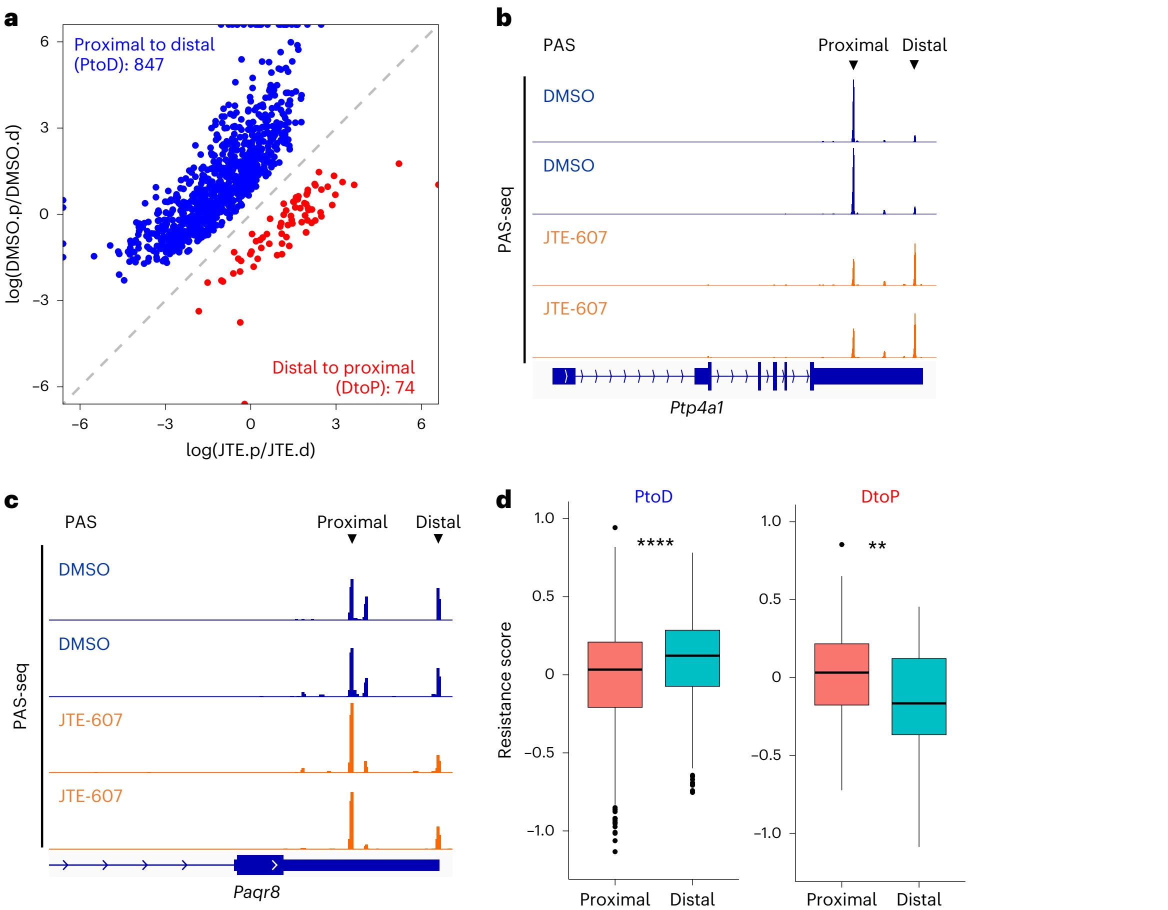


Fig. 5 | JTE-607-induced APA changes in cells are sequence specific. a, A scatter plot showing JTE-607-induced APA changes in cells. **b**,**c**, PAS-seq tracks of two example genes: *Ptp4a1* (**b**) and *Paqr8* (**c**) ($n = 2$ biological replicates and the positions of the proximal and distal PASs are marked). **d**, Boxplots comparing the C3PO-predicted resistance scores for the proximal (Prox) and distal (Dist) PASs of the Prox-to-Dist (847 genes) and Dist-to-Prox (74 genes) genes. $^{****}P < 0.0001$; $^{*}P < 0.05$ (two-sided Student's *t*-test). Cohen's *d*: Prox-to-Dist: −0.38; Dist-to-Prox: 0.47. For all boxplots, box limits give the interquartile range with whiskers extending by a factor of 1.5 and the center line showing the median.

sensitivities of PASs through two analyses. First, we selected genes with the top 1,000 resistant or sensitive PASs based on the C3PO-predicted resistance scores. To avoid complications from neighboring genes, we selected genes that do not overlap with other genes in the 1-kb downstream region for our analyses. The average normalized 4sU-seq signals at genes with the top 1,000 resistant PASs showed that transcription terminated efficiently at these PASs in both DMSO- and JTE-607-treated cells, and an increase in RT levels was observed downstream of the PASs (Fig. 6c, red arrow). By contrast, for genes with the top 1,000 sensitive PASs, their global 4sU-seq signals revealed significantly higher RT in JTE-607-treated cells compared with DMSO-treated cells (Fig. 6d, red arrow), suggesting that JTE-607 induced significant inhibition of mRNA 3′ processing at these PASs. The JTE-607-induced increase in RT levels was significantly higher at the sensitive PASs than the resistant sites (Fig. 6e, $P < 2.2 \times 10^{-16}$, Wilcoxon's test). Second, based on our 4sU-seq data, we identified 1,000 genes with the most significant RT and another 1,000 with the least RT using a computational tool called ARTDeco[28]. When comparing the PASs of these gene groups, we observed that the genes with the lowest JTE-607-induced RT displayed significantly higher C3PO-predicted resistance scores than those with high RT (Extended Data Fig. 8). Together, our PAS-seq and 4sU-seq analyses suggest that JTE-607 inhibits mRNA 3′ processing and transcription termination in a sequence-dependent manner and that C3PO-predicted drug sensitivity can provide a basis for understanding the global effect of JTE-607 on PAS selection and transcription termination.

We also wanted to compare the effect of JTE-607-mediated inhibition of CPSF73 with that of CPSF73 depletion. Liu et al. recently showed that UBE3D is required for CPSF73 stability and *UBE3D* knockdown led to efficient CPSF73 depletion[20]. Therefore, we compared the APA changes caused by JTE-607 and those caused by *UBE3D* knockdown. Our analysis showed that *UBE3D* depletion led to widespread APA changes, characterized primarily by 3′-UTR lengthening (Extended Data Fig. 9a), which is consistent with other studies of CPSF73 knockdown[29]. *UBE3D* knockdown and JTE-607 resulted in APA changes in overlapping as well as distinct sets of genes (Extended Data Fig. 9b). These results indicate that JTE-607-mediated inhibition of CPSF73 and CPSF73 depletion impact mRNA 3′ processing through different mechanisms.

Nucleotide composition of the resistant and sensitive human PASs revealed distinct patterns. JTE-607-resistant PASs have alternating U- and A-rich regions (Fig. 6f, left) whereas the JTE-607-sensitive PASs are generally U/G rich (Fig. 6f, middle). These patterns are very consistent with the top resistant and sensitive PASs from our MPIVA screen (Fig. 3f–g). It is interesting that the average nucleotide composition of the CS regions of all annotated human PASs also displayed alternating U- and A-rich regions (Fig. 6f, right), suggesting that a portion of the human PASs is potentially resistant to JTE-607. Furthermore, a comparison of the resistant and sensitive PASs revealed that the resistant PASs are more conserved than the sensitive PASs (Extended Data Fig. 10a). Finally, gene ontology analysis of the genes expressed in HepG2 cells

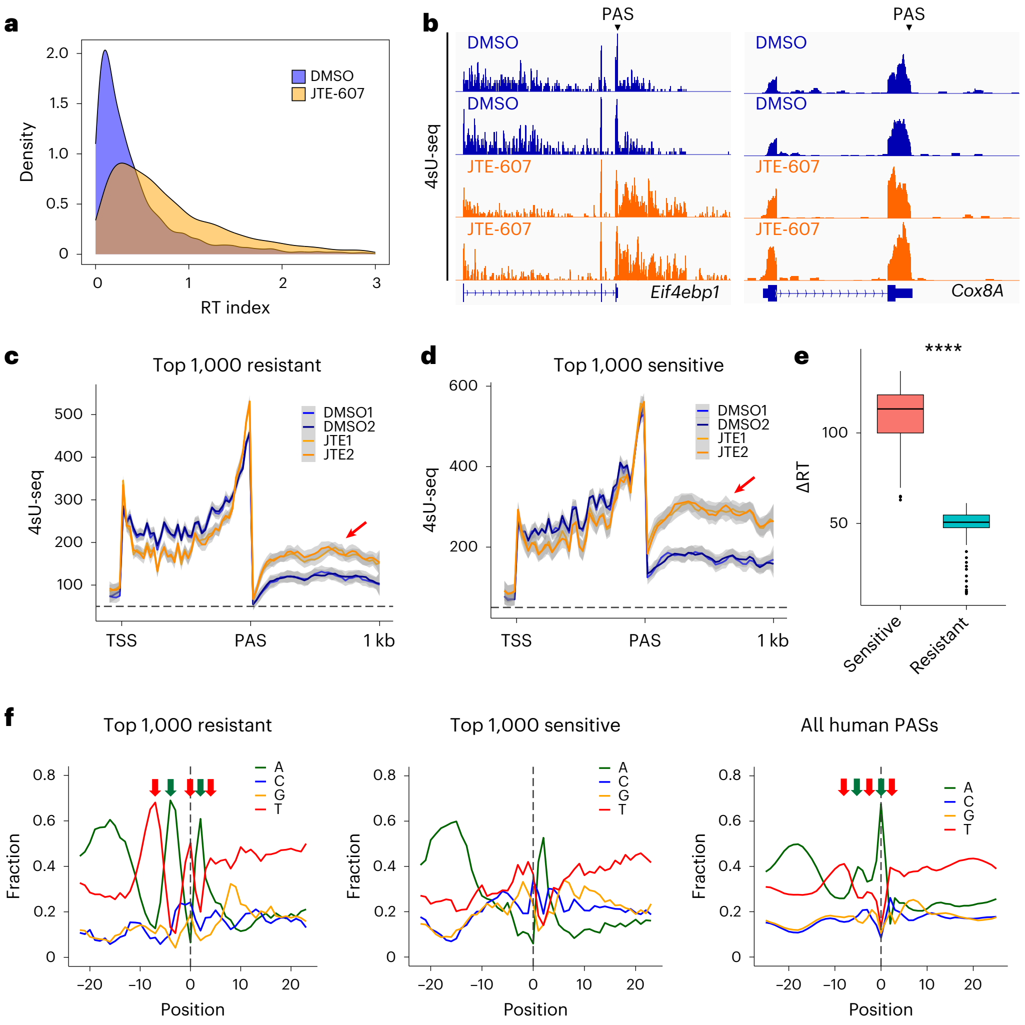


Fig. 6 | JTE-607-mediated inhibition of mRNA 3′ processing in cells is sequence specific. a, A density plot of transcription RT index (read counts in the 1-kb downstream region per read counts in gene body) for DMSO- and JTE-607-treated cells based on 4sU-seq data. **b**, The 4sU-seq tracks for *Eif4ebp1* and *Cox8A* genes (*n* = 2 biological replicates for DMSO- and JTE-607-treated cells). PAS positions are marked. **c**, Average normalized 4sU-seq signals for the genes with the top 1,000 most resistant PASs. **d**, Similar to **c**, but for the top 1,000 most sensitive PASs. The red arrow denotes the region downstream of PAS. TSS, Transcription start site. **e**, A box plot comparing the ΔRT (the difference in 4sU-seq signals in the 1-kb region downstream of the PAS shown in **c** and **d**). ****$P < 0.0001$, two-sided Wilcoxon's test. **f**, CS region nucleotide distribution for the top 1,000 most resistant (left), most sensitive (middle) and all human (right) PASs. The T- and A-rich regions are marked by red and green arrows.

containing the top 1,000 sensitive and resistant PASs revealed different enriched functional categories (Extended Data Fig. 10b).

Discussion

In the present study, we characterized JTE-607, a new inhibitor of the endonuclease for mRNA 3′ processing, CPSF73. Although CPSF73 is universally required for mRNA 3′ processing, we have unexpectedly discovered that compound 2 inhibits the cleavage step of mRNA 3′ processing in a sequence-dependent manner both in vitro and in cells, and that the CS region sequence is a major determinant of drug sensitivity. We have characterized the relationship between the CS region sequence and compound 2 sensitivity using MPIVA coupled with machine learning. Our machine learning model C3PO can predict compound 2 sensitivity based on CS sequence and provides the basis for understanding the impact of JTE-607 on APA and transcription termination in human cells. Therefore, our study not only has provided new insights into the fundamental mechanism of mRNA 3′ processing, but may also have important implications for the use of JTE-607 as a research and therapeutic tool.

What is the molecular mechanism for the sequence-specific sensitivity to compound 2? As both compound 2 and the CS region RNA bind to CPSF73 at or near its active site[18,30], these interactions are competitive and most likely mutually exclusive (Fig. 7). As the CPSF73 affinity for compound 2 is constant, the outcome of this competition is thus determined by the affinity of the CPSF73–CS region RNA interaction. Based on the structure of the histone mRNA-cleavage complex[30], which contains the nuclease module consisting of CPSF73, CPSF100 and symplekin, these proteins form an RNA-binding channel that can bind to an ~20-nt CS sequence (Fig. 7). Other mRNA 3′ processing factors can also be involved, including Fip1 and PAP. Fip1 is known to bind to U-rich sequences near the AAUAAA hexamer[31,32] and the compound 2-resistant CS sequences contain U-rich sequences (Fig. 3f,g). Finally, an early study showed that PAP is required for in vitro cleavage of L3 PAS, but not for SVL, and that the CS region sequences determine the PAP dependency[33]. Given the important roles for the CS region in determining both PAP dependency and compound 2 sensitivity, it is possible that PAP is involved in binding to CS-region sequences. Based on these results, we propose that CPSF73 and other mRNA 3′ processing factors

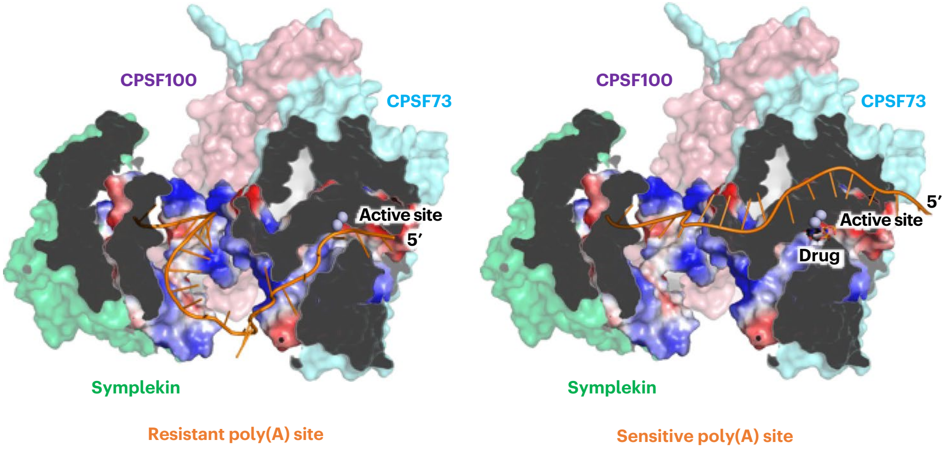


Fig. 7 | A model for sequence-specific inhibition of pre-mRNA 3′ processing by compound 2. Artistic rendering of resistant (left) or sensitive (right) PAS RNAs within the pre-mRNA 3′ processing complex. CPSF73, CPSF100 and symplekin are shown and their structures are partially based on the histone mRNA-cleavage complex (PDB accession no. 6V4X) and the RNA-binding channel formed by the three proteins is highlighted. The RNA (orange thread), active site of CPSF73 and the drug (compound 2) are marked. Please see Discussion for details.

form an RNA-binding channel that directly binds to the CS region in a sequence-specific manner, and that the affinity of this interaction determines the compound 2 sensitivity of a PAS (Fig. 7).

The nucleotide composition in the CS region has been conserved from yeast to human[9–11] and this pattern is highly similar to that of the compound 2-resistant PASs (Fig. 6f). In addition, compound 2-resistant PASs are more evolutionarily conserved than the sensitive sites (Extended Data Fig. 10a). It remains unclear what, if any, selection pressure favors PASs that are resistant to a small molecule, that is, not present in most environments. One possibility is that the CS region sequence may impact transcription termination. According to our model, the resistant PASs interact with CPSF73 and other mRNA 3′ processing factors more strongly (Fig. 7). As the mRNA 3′ processing machinery is known to directly bind to RNA polymerase II[34–36], such interaction could contribute to slowing down of the polymerase, thereby promoting termination. Thus, a subset of PASs may have evolved to promote transcription termination by binding to CPSF73 more strongly via their CS and the compound 2 resistance is an unintended consequence.

Our results may have implications for understanding how JTE-607 specifically kills myeloid leukemia and Ewing's sarcoma cell lines. As mentioned earlier, JTE-607 is a prodrug and is converted to compound 2 by the cellular enzyme CES1 (ref. 18). Although cellular CES1 levels may contribute to the cell-type specificity, previous studies showed that CES1 level is a poor predictor for JTE-607 sensitivity[18]. Thus, the molecular basis for cell-type-specific toxicity of JTE-607 remains unknown. Based on the results reported in the present study, we propose two possible mechanisms for explaining the cell-type-specific drug sensitivity. First, the potency for JTE-607-mediated inhibition of mRNA 3′ processing may be cell-type specific. Our model suggests that the drug sensitivity is determined by the interaction affinity between the CPSF73 and other mRNA 3′ processing factors and the CS region sequence. If cell-type-specific mechanisms can modulate the specificity of this interaction, they can alter JTE-607 sensitivity globally. This could result from cell-type-specific expression levels or posttranslational modification of CPSF73 and other mRNA 3′ processing factors that bind to the CS region. A recent study has provided support for this model[37]. Alternatively, the sequence specificity of JTE-607 is similar among different cell types. However, cells of myeloid leukemia and Ewing's sarcoma may be uniquely dependent on one gene or a subset of genes with PASs that are highly sensitive to JTE-607. For example, a recent study identified PDXK, an enzyme in the vitamin B_6 metabolism pathway, as a unique acute myeloid leukemia dependency gene[38]. If the PASs of such dependency genes are sensitive to JTE-607, the expression of these genes would be repressed by JTE-607 treatment, leading to cell death in specific cell types. Further studies are needed to distinguish between these models.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41594-023-01161-x.

References

1. Colgan, D. F. & Manley, J. L. Mechanism and regulation of mRNA polyadenylation. *Genes Dev.* **11**, 2755–2766 (1997).
2. Chan, S., Choi, E. A. & Shi, Y. Pre-mRNA 3′-end processing complex assembly and function. *Wiley Interdiscip. Rev. RNA* **2**, 321–335 (2011).
3. Shi, Y. Alternative polyadenylation: new insights from global analyses. *RNA* **18**, 2105–2117 (2012).
4. Mitschka, S. & Mayr, C. Context-specific regulation and function of mRNA alternative polyadenylation. *Nat. Rev. Mol. Cell Biol.* https://doi.org/10.1038/s41580-022-00507-5 (2022).
5. Tian, B. & Manley, J. L. Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.* **18**, 18–30 (2016).
6. Mandel, C. R. et al. Polyadenylation factor CPSF-73 is the pre-mRNA 3′-end-processing endonuclease. *Nature* **444**, 953–956 (2006).
7. Shi, Y. & Manley, J. L. The end of the message: multiple protein–RNA interactions define the mRNA polyadenylation site. *Genes Dev.* **29**, 889–897 (2015).
8. Sheets, M. D., Ogg, S. C. & Wickens, M. P. Point mutations in AAUAAA and the poly(A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Res.* **18**, 5799–5805 (1990).
9. Ozsolak, F. et al. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* **143**, 1018–1029 (2010).
10. Liu, X. et al. Comparative analysis of alternative polyadenylation in *S. cerevisiae* and *S. pombe*. *Genome Res.* **27**, 1685–1695 (2017).
11. Derti, A. et al. A quantitative atlas of polyadenylation in five mammals. *Genome Res.* **22**, 1173–1183 (2012).

12. Palencia, A. et al. Targeting *Toxoplasma gondii* CPSF3 as a new approach to control toxoplasmosis. *EMBO Mol. Med.* **9**, 385–394 (2017).
13. Begolo, D. et al. The trypanocidal benzoxaborole AN7973 inhibits trypanosome mRNA processing. *PLoS Pathog.* **14**, e1007315 (2018).
14. Sonoiki, E. et al. A potent antimalarial benzoxaborole targets a *Plasmodium falciparum* cleavage and polyadenylation specificity factor homologue. *Nat. Commun.* **8**, 1–11 (2017).
15. Sasaki, J. et al. Prior burn insult induces lethal acute lung injury in endotoxemic mice: effects of cytokine inhibition. *Am. J. Physiol. Lung Cell Mol. Physiol.* **284**, L270–L278 (2003).
16. Uesato, N., Fukui, K., Maruhashi, J., Tojo, A. & Tajima, N. JTE-607, a multiple cytokine production inhibitor, ameliorates disease in a SCID mouse xenograft acute myeloid leukemia model. *Exp. Hematol.* **34**, 1385–1392 (2006).
17. Jian, M. Y., Koizumi, T., Tsushima, K. & Kubo, K. JTE-607, a cytokine release blocker, attenuates acid aspiration-induced lung injury in rats. *Eur. J. Pharmacol.* **488**, 231–238 (2004).
18. Ross, N. T. et al. CPSF3-dependent pre-mRNA processing as a druggable node in AML and Ewing's sarcoma. *Nat. Chem. Biol.* **16**, 50–59 (2020).
19. Kakegawa, J., Sakane, N., Suzuki, K. & Yoshida, T. JTE-607, a multiple cytokine production inhibitor, targets CPSF3 and inhibits pre-mRNA processing. *Biochem. Biophys. Res. Commun.* **518**, 32–37 (2019).
20. Liu, H., Heller-Trulli, D. & Moore, C. L. Targeting the mRNA endonuclease CPSF73 inhibits breast cancer cell migration, invasion, and self-renewal. *iScience* **25**, 104804 (2022).
21. Boreikaite, V., Elliott, T. S., Chin, J. W. & Passmore, L. A. RBBP6 activates the pre-mRNA 3′ end processing machinery in humans. *Genes Dev.* **36**, 210–224 (2022).
22. Gutierrez, P. A., Baughman, K., Sun, Y. & Tong, L. A real-time fluorescence assay for CPSF73, the nuclease for pre-mRNA 3′-end processing. *RNA* **27**, 1148–1154 (2021).
23. Sample, P. J. et al. Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. *Nat. Biotechnol.* **37**, 803–809 (2019).
24. Wu, X. & Bartel, D. P. Widespread influence of 3′-end structures on mammalian mRNA processing and stability. *Cell* **169**, 905–917.e11 (2017).
25. Linder, J., Koplik, S. E., Kundaje, A. & Seelig, G. Deciphering the impact of genetic variation on human polyadenylation using APARENT2. *Genome Biol.* **23**, 232 (2022).
26. Yoon, Y., Soles, L. V. & Shi, Y. PAS-seq 2: a fast and sensitive method for global profiling of polyadenylated RNAs. *Methods Enzymol.* **655**, 25–35 (2021).
27. Nojima, T. et al. Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* **161**, 526–540 (2015).
28. Roth, S. J., Heinz, S. & Benner, C. ARTDeco: automatic readthrough transcription detection. *BMC Bioinform.* **21**, 214 (2020).
29. Li, W. et al. Systematic profiling of polyA$^+$ transcripts modulated by core 3′ end processing and splicing factors reveals regulatory rules of alternative cleavage and polyadenylation. *PLoS Genet.* **11**, 1–28 (2015).
30. Sun, Y. et al. Structure of an active human histone pre-mRNA 3′-end processing machinery. *Science* **367**, 700–703 (2020).
31. Lackford, B. et al. Fip1 regulates mRNA alternative polyadenylation to promote stem cell self-renewal. *EMBO J.* **33**, 878–889 (2014).
32. Martin, G., Gruber, A. R., Keller, W. & Zavolan, M. Genome-wide znalysis of pre-mRNA 3′ end processing reveals a decisive role of human cleavage factor I in the regulation of 3′ UTR length. *Cell Rep.* **1**, 753–763 (2012).
33. Ryner, L. C., Takagaki, Y. & Manley, J. L. Sequences downstream of AAUAAA signals affect pre-mRNA cleavage and polyadenylation in vitro both directly and indirectly. *Mol. Cell Biol.* **9**, 1759–1771 (1989).
34. Richard, P. & Manley, J. L. Transcription termination by nuclear RNA polymerases. *Genes Dev.* **23**, 1247–1269 (2009).
35. Bentley, D. L. Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. *Curr. Opin. Cell Biol.* **17**, 251–256 (2005).
36. Proudfoot, N. J. Transcriptional termination in mammals: stopping the RNA polymerase II juggernaut. *Science* **352**, aad9926 (2016).
37. Cui, Y. et al. Elevated pre-mRNA 3′ end processing activity in cancer cells renders vulnerability to inhibition of cleavage and polyadenylation. *Nat. Commun.* **14**, 4480 (2023).
38. Chen, C.-C. et al. Vitamin B6 addiction in acute myeloid leukemia. *Cancer Cell* **37**, 71–84.e7 (2020).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Cell culture

HepG2 cells were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% (v:v) fetal bovine serum (FBS) at 37 °C in a 5% (v:v) CO_2-enriched incubator. HeLa S3 cells were maintained in MEM Joklik Modification supplemented with 2.4 mM sodium bicarbonate and 8% (v:v) newborn calf serum in a spinner flask at 37 °C with ambient CO_2. For JTE-607 treatment, 20 µM final concentration of JTE-607 (Tocris) in DMSO was added to the cell culture medium and incubated at 37 °C for 4 h.

In vitro mRNA 3′ processing assay

HeLa NE was made as previously described[39]. All PASs were cloned into the pBlueScript II KS+ vector. RNA substrates were synthesized by run-off in vitro transcription (IVT) using T7 polymerase (New England Biolabs (NEB)) in the presence of [α-^{32}P]UTP according to the manufacture's protocol. For in vitro cleavage reaction with compound 2, NE was pre-incubated with 10% DMSO or compound 2 in 10% DMSO for 30 min on ice. Each 10 µl reaction contains 20 c.p.s. (counts per s) of radiolabeled RNA, 44% (v:v) NE, 8.8 mM Hepes, pH 7.9, 44 mM KCl, 0.44 mM $MgCl_2$, 0.2 mM 3′-dATP (Sigma-Aldrich), 2.5% (v:v) polyvinyl alcohol (PVA), 40 mM creatine phosphate and 4 mM 2-mercaptoethanol. The reaction mix was incubated for 90 min at 30 °C. RNA was extracted, resolved on an 8% urea–polyacrylamide gel electrophoresis (PAGE) and visualized by phosphor imaging. The IC_{50} was calculated using the equation: [Inhibitor] versus normalized response – Variable slope on Prism.

EMSA

In vitro cleavage reactions, 10 µl, without PVA were incubated for 20 min at 30 °C, heparin was added to 0.4 µg µl^{-1} and the reaction was incubated for an additional 5 min on ice and resolved on 4% native PAGE in 1× Tris-glycine running buffer, pH 8.3 at 100 V for 4 h in an ice bath. The gel was dried and visualized by phosphor imaging.

MPIVA

DNA oligos containing 23 random nucleotide CS was purchased from IDT and PCR amplified to generate double-stranded (ds)DNA. The dsDNA library was cloned into pBlueScript II KS+ vector by Gibson Assembly (NEB) and electroporated into ElectroMAX DH5α (Thermo Fisher Scientific). Library quality control was determined as previously described[40]. RNA pools were synthesized by IVT using T7 polymerase (NEB) followed by RQ1 DNase treatment (Promega). The RNA pool was purified by phenol–chloroform extraction and was either polyadenylated (for input) or 3′-dATP blocked (for DMSO and compound 2 treated) by *Escherichia coli* PAP (NEB). The RNAs were incubated in 600-µl reactions containing 6 pmol of RNA, 44% (v:v) HeLa NE, 8.8 mM Hepes-OH, pH 7.9, 44 mM KCl, 1.44 mM $MgCl_2$, 1 mM ATP, 2.5% (v:v) PVA, 20 mM creatine phosphate, 4 mM 2-mercaptoethanol and either 1% DMSO or compound 2 in DMSO. The reaction mixture was incubated for 90 min at 30 °C. Polyadenylated RNAs were isolated using NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB) and reverse transcribed using SuperScript III reverse transcriptase (Invitrogen) with an anchored oligo dT primer. The complementary DNA library was bead purified (Beckman Coulter) and amplified using a library-specific forward primer and reverse primer containing Illumina adapter sequences and a linker. The amplified libraries were resolved on a 2.5% low melting point agarose gel and extracted.

All MPIVA read 1 and 2 FASTQ files were merged using bbmerge v.38 with 'maxloose = t'[41]. Untreated RNA sequencing (RNA-seq) reads established the sequences of the full randomized PAS region contained in the IVT pool. Then 25-nt randomized regions were clustered using starcode v.1.4 (ref. 42) to account for sequencing errors and determine consensus sequences. The expected cleaved lengths of the 25-nt consensus sequences for L3 and SVL backbones (13 nt and 12 nt, respectively) were used as unique identifiers of the full randomized region. If any identifiers were not unique within L3 and SVL libraries, then these sequences were removed from subsequent analyses.

Next, RNA-seq reads from DMSO- and drug-treated libraries were locally aligned against shared reporter regions and PAS consensus sequences. Only reads with a unique consensus sequence alignment were kept and used to determine cut sites. Additional checks and corrections were performed to ensure that cut sites were not mis-assigned inside the poly(A) tail. Then, 158,298 L3 variants and 103,018 SVL variants were left after filtering for sequences with at least 50 reads in the DMSO libraries. A pseudocount of 1 was then added to variants in L3 2.5 µM, L3 12.5 µM and SVL 12.5 µM owing to drug-mediated dropout of variants in the DMSO libraries. This pseudocount avoids having undefined drug sensitivities resulting from log(0). Each variant that passed these checks was counted and converted to a percentage by dividing by the total number of kept reads. Drug sensitivity for each variant in each dose of compound 2 was defined as the log(ratio) of normalized reads from drug-treated RNA-seq divided by the normalized reads from DMSO-treated RNA-seq. Within a given drug dose, sequences with higher log(ratios) are more resistant than those sequences with lower log(ratios).

MFE folding of IVT RNAs. MFE predictions were done with RNAStructure v.6.4's *Fold*[43] using the entire IVT RNA sequence. Δ*G* values of each MFE were determined with RNAStructure's *efn2* with argument '--simple'. Δ*G* values from the top 10,000 resistant and sensitive sequences were compared ('Quantification and statistical analysis').

C3PO machine learning architecture and training. The architecture chosen for predicting drug sensitivity is based on a previously published three-layer CNN for predicting polysome profiles[23]. The model takes in 25-nt one-hot-encoded sequences followed by:

- First convolution layer: 120 filters (8 × 4), batch normalization, ReLU (rectified linear unit) activation, zero padding to maintain the same length input and output and 0% dropout.
- Second convolution layer: 120 filters (8 × 1), batch normalization, ReLU activation, zero padding to maintain the same length input and output and 0% dropout.
- Third convolution layer: 120 filters (8 × 1), batch normalization, ReLU activation, zero padding to maintain the same length input and output and 0% dropout.
- Dense layer: 80 nodes, batch normalization, ReLU activation and 10% dropout.
- Output layer: three linear outputs.

The Adam optimizer[44] was used for model fitting with a mean squared error loss function, batch size of 64 and sample weights based on the DMSO read depth.

Sequences were assembled into test and training sets to mix highly covered variants from both RNA contexts (L3 and SVL) into the test and training sets. Within each RNA context, sequences were ordered by DMSO read depth, split based on the sequences' number in this ordering into odd and even lists and then concatenated. Finally, the L3 and SVL sequences were interleaved to make an even coverage between RNA contexts in the test set. The test set consisted of the top 4,120 sequences and the training set the remaining sequences. The test set size was chosen because it reflects 2% of the variant space in SVL which contains fewer variants than L3.

Ten iterations of training with six epochs were conducted to account for slight variations in model performance resulting from training algorithm stochasticity. Performance between iterations was evaluated by the square of Pearson's *r* (R^2) between measured and predicted compound 2 sensitivity in the test sequences. The best performing iteration was kept and used in further analyses.

Exploring additional machine learning architectures and training. Additional machine learning and training pipelines were explored based on CNNs and dilated residual networks. With the C3PO CNN architecture, training was done with four to eight epochs and this number of epochs performed relatively similarly on the test set (Supplementary Table 2). We also explored using a validation set (4,120 sequences) derived from the training set to determine an early stopping criterion for the number of epochs trained, and this performed similarly to the models trained with a preset number of epochs. Among the three drug doses predictions, 12.5-µM predictions performed better, leading us to train models for only this dose. However, 12.5-µM prediction performance between the three- and one-dose predictions was negligible, so we used the model with three dose predictions.

Hyperband training[45] with the CNN architecture was also performed to ascertain potential optimal hyperparameter values. Hyperparameters were allowed to range from 1 to 5 on-dimensional (1D) convolutional layers with ReLU activation and batch normalization, 8–140 (step 16) number of filters, followed by pooling choices of average, maximum or none and dropout rates of 0–0.5 (step 0.1). These convolutional layer(s) are followed by a Flatten layer and one to three dense layers. Each dense layer can be of size 20–200 (step 20) with ReLU activation, batch normalization and dropout rates of 0–0.5 (step 0.1). Learning rate parameters were also allowed to range between 1×10^{-5} and 1×10^{-1}. Training was allowed to stop early based on the validation set's mean squared error and a minimum delta of 0.001 and patience of five epochs. Hyperband training was done with an output layer for all three drug doses and for predicting only 12.5-µM compound 2 resistance.

We tested the residual neural network (ResNet) architecture with predicting both compound 2 sensitivity and cleavage patterns with the hypothesis that learning sequence features that affect CS usage would improve the compound 2 sensitivity predictions. Input to the ResNet is a one-hot-encoded, 25-nt sequence and is followed by 20 residual blocks, where each block contains 2 layers of dilated convolutions and a skip connection. More specifically, there are 5 residual groups where each residual group contains 4 residual blocks with 32 channels and convolutional filters of size 3. Each residual block is encoded the same as APARENT2 (ref. 25) where each block has two 1D convolutional layers with batch normalization, ReLU activation and a filter dilatation rate. There are additional skip connections from between each residual group and the last convolutional layer and produce a vector of length 26, $s(x)$. The 26th position is for all cuts outside of the 25-nt randomized region. For training and accounting for any background sequence biases, a boolean is passed to indicate whether the data point is from the L3 or SVL background, which is multiplied with a position-specific weight matrix and linearly combined with $s(x)$. We also kept APARENT2's random shifting of the input sequence and cleavage distribution during training to force the network to not simply learn the designed expected cleavage position in each library. These scores containing library-specific information are sent to four different linear dense layers for separate predictions of cleavage profiles of all four drug doses and softmax transformation is applied to each. For compound 2 sensitivity prediction, $s(x)$ undergoes average pooling and the library indicator is concatenated before a linear dense layer for final output. Kullback–Leibler divergence is used as the loss function for cleavage profiles and mean squared error for compound 2 sensitivities. Total loss is a weighted average of half from compound 2 sensitivities and the other half split evenly between the four cleavage profiles. The ResNet was trained with Keras's implementation of the Adam optimizer, batch size of 64 and stopping criteria based on a validation set (4,120 sequences) derived from the training set.

We first tried 1, 2, 4, 2 and 1 as dilatation rates for the 5 residual groups and performed similarly to previously trained CNNs but did not outperform C3PO (Supplementary Table 2). We also tried lower dilatation rates of 1, 2, 2, 2 and 1 as well as 1, 1, 1, 1 and 1, which performed worse. Using the dilatation rates 1, 2, 4, 2 and 1, we trained for exactly 7 epochs and did not find improved performance. We also increased the cleavage profile length to 27 to separately model cuts found at positions greater than the 25-nt randomized region in position 26, and position 27 is filled when a sequence is not found at a given compound 2 dose (that is, sensitive sequences that drop out at higher compound 2 doses). Finally, we increased the weight of compound 2 sensitivity predictions to 75% of the total loss. These ResNet variations did not lead to a better performance than C3PO (Supplementary Table 2).

Convolutional layers 1 and 2 activation analysis. Convolutional layers 1 and 2 were analyzed similarly to a previously published analysis of a CNN (APARENT)[40]. In brief, every filter in both convolutional layers was correlated with predictions of drug sensitivity at the 12.5-µM dose. The top 5,000 input sequences from the training set that achieved maximal filter activation were put into a position weight matrix and used to generate position-aware consensus sequence logos[46]. Pearson's r plots of each filter's activations with predicted 12.5 µM compound 2 sensitivity at each position are plotted below these filter-specific sequence logos. Layer 1 filters are 8 positions wide and layer 2 filters are 15 positions wide. Note that the convolutional layers in C3PO contain even zero padding to maintain an input/output size of 25.

APARENT2 predictions and comparisons. APARENT2 predictions of log(odds) of cleavage at the expected cleavage position versus elsewhere were done on all MPIVA sequences, centered at their expected cut site. Predictions with read depth of at least 150 in the Input libraries were kept for further analysis. APARENT2 predictions were compared against the log(odds) of expected cleaved DMSO read counts and input read counts, which estimates the in vitro cleavage efficiency. In addition, APARENT2 predictions were compared against the log(odds) of expected cleaved 12.5 µM compound 2 and input read counts which estimates the in vitro drug resistance.

Motif analysis. The top 10,000 resistant and sensitive sequences were converted into their 6-mer counts and significant 6-mers were determined ('Quantification and statistical analysis'). The nucleotide content of significant resistant and sensitive 6-mers are shown next to their respective axes (Supplementary Fig. 6).

4sU-seq

HepG2 cells were treated with DMSO or 20 µM JTE-607 (Tocris) for 3 h at 37 °C. Then, 500 µM 4sU (Sigma-Aldrich) was added to the DMSO-/JTE-607 containing medium and cells were incubated at 37 °C for an additional 1 h. Cells were lysed in TRIzol (Invitrogen) and total RNA was extracted following the manufacturer's protocol. The 4sU RNA enrichment and library preparation were done as previously described[47]. All sequencing reads were mapped to human hg19 using STAR[48].

PAS-seq

HepG2 cells were treated with DMSO or 20 µM JTE-607 (Tocris) for 4 h at 37 °C and total RNA was extracted by TRIzol. PAS-seq library preparation and data analyses were performed as previously described[49].

Quantification and statistical analysis

Pearson's r and R^2 (square of Pearson's r) are used in Figs. 1–4, Extended Data Figs. 2, 3, 6 and 7 and related text, as well as Supplementary Models Table. Potential inequality of the top 10,000 resistant and sensitive sequences' MFE ΔG values were tested using a two-sided Student's t-test with unequal variance. The 6-mers in the top 10,000 resistant and sensitive sequences were found to be significant by a binomial test with a null hypothesis of probability of success = 0.25^6 and alternative hypothesis of $>0.25^6$. The P-value threshold was adjusted by the number of possible k-mers, 4^6, and thus significant 6-mers must have P values $\leq \frac{0.05}{4^6}$. Cohen's d and Hedge's g calculations were used to determine the effect size of the difference between two group means and calculated using the Python package pingouin[50].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All deep sequencing data from the present study have been deposited in the Gene Expression Omnibus under series accession no. GSE218977. Source data are provided with this paper.

Code availability

The codes for machine learning analysis from the present study have been deposited to GitHub at https://github.com/angelamyu/C3PO.

References

39. Abmayr, S. M., Yao, T., Parmely, T. & Workman, J. L. Preparation of nuclear and cytoplasmic extracts from mammalian cells. *Curr. Protoc. Mol. Biol.* **75**, 12.1.1–12.1.10 (2006).
40. Bogard, N., Linder, J., Rosenberg, A. B. & Seelig, G. A deep neural network for predicting and engineering alternative polyadenylation. *Cell* **178**, 91–106.e23 (2019).
41. Bushnell, B., Rood, J. & Singer, E. BBMerge—accurate paired shotgun read merging via overlap. *PLoS ONE* **12**, e0185056 (2017).
42. Zorita, E., Cuscó, P. & Filion, G. J. Starcode: sequence clustering based on all-pairs search. *Bioinformatics* **31**, 1913–1919 (2015).
43. Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinform.* **11**, 129 (2010).
44. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. Preprint at http://arxiv.org/abs/1412.6980 (2017).
45. Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A. & Talwalkar, A. Hyperband: a novel bandit-based approach to hyperparameter optimization. *J. Mach. Learn. Res.* **18**, 6765–6816 (2017).
46. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
47. Wang, X. et al. Herpes simplex virus blocks host transcription termination via the bimodal activities of ICP27. *Nat. Commun.* **11**, 293 (2020).
48. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
49. Wang, X. et al. Mechanism and consequences of herpes simplex virus 1-mediated regulation of host mRNA alternative polyadenylation. *PLoS Genet.* **17**, e1009263 (2021).
50. Vallat, R. Pingouin: statistics in Python. *J. Open Source Softw.* **3**, 1026 (2018).

Acknowledgements

We thank R. Beckwith for providing compound 2 and B. Semler and J. H. C. Nguyen for providing the suspension HeLa cell line. We wish to acknowledge the support of the Chao Family Comprehensive Cancer Center Shared Resource Genomics High-Throughput Facility, supported by the National Cancer Institute of the National Institutes of Health (NIH) under award no. P30CA062203. The present study was supported by the following grants: NIH (grant nos. R01AI166703, R01GM090056 and R35GM149294 to Y.S and R21HG010945 to G.S). L.L. is supported by the Center for Virus Research Graduate Fellowship provided by the University of California, Irvine Graduate Division. A.M.Y. is a Washington Research Foundation Postdoctoral Fellow. R.C.S. is a Pew Biomedical Scholar.

Author contributions

L.L., A.M.Y., G.S. and Y.S. conceived and designed the project. L.L. and X.W. performed all the biochemical and sequencing experiments with help from L.V.S., Y.Y., K.S.K.S. and M.C.V. A.M.Y. performed all the machine learning experiments and data analysis. X.T., Y.C., J.L., W.E., R.S., Z.Y., I.M., F.Q., W.L. and Y.S. performed all the other bioinformatic analysis. L.L., A.M.Y., G.S. and Y.S. wrote the paper with input from all authors.

Competing interests

The authors declare no competing interests.

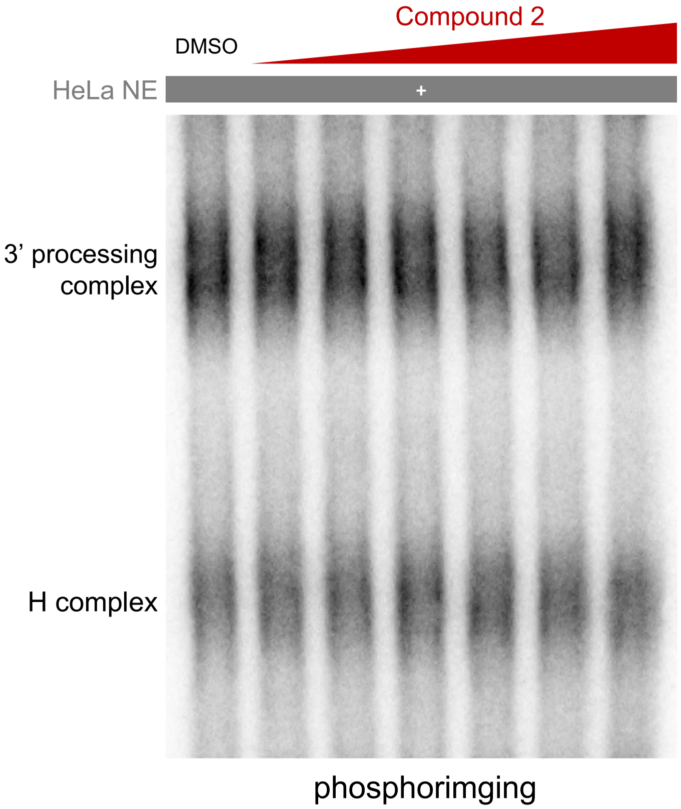
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41594-023-01161-x.

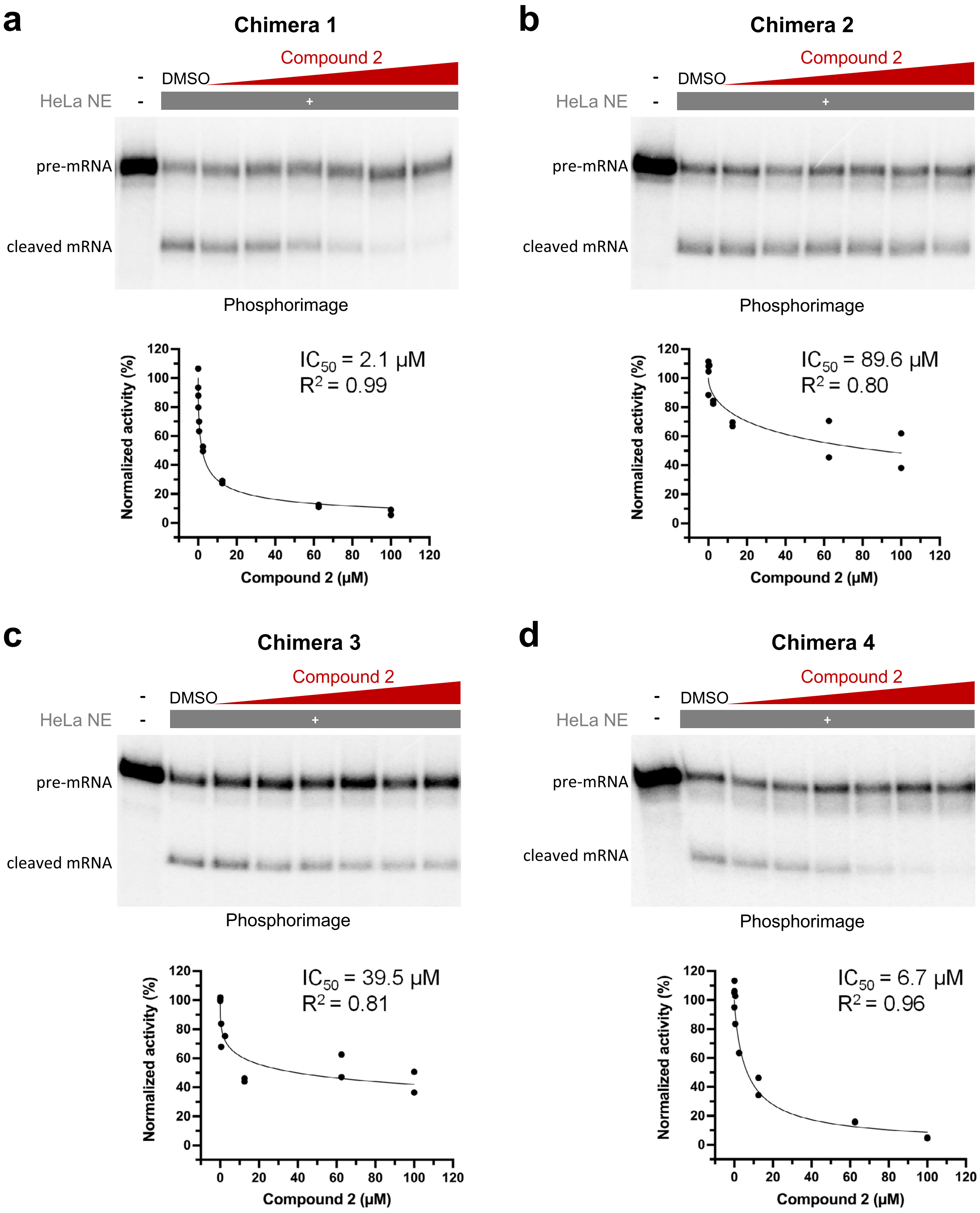
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41594-023-01161-x.

Correspondence and requests for materials should be addressed to Georg Seelig or Yongsheng Shi.

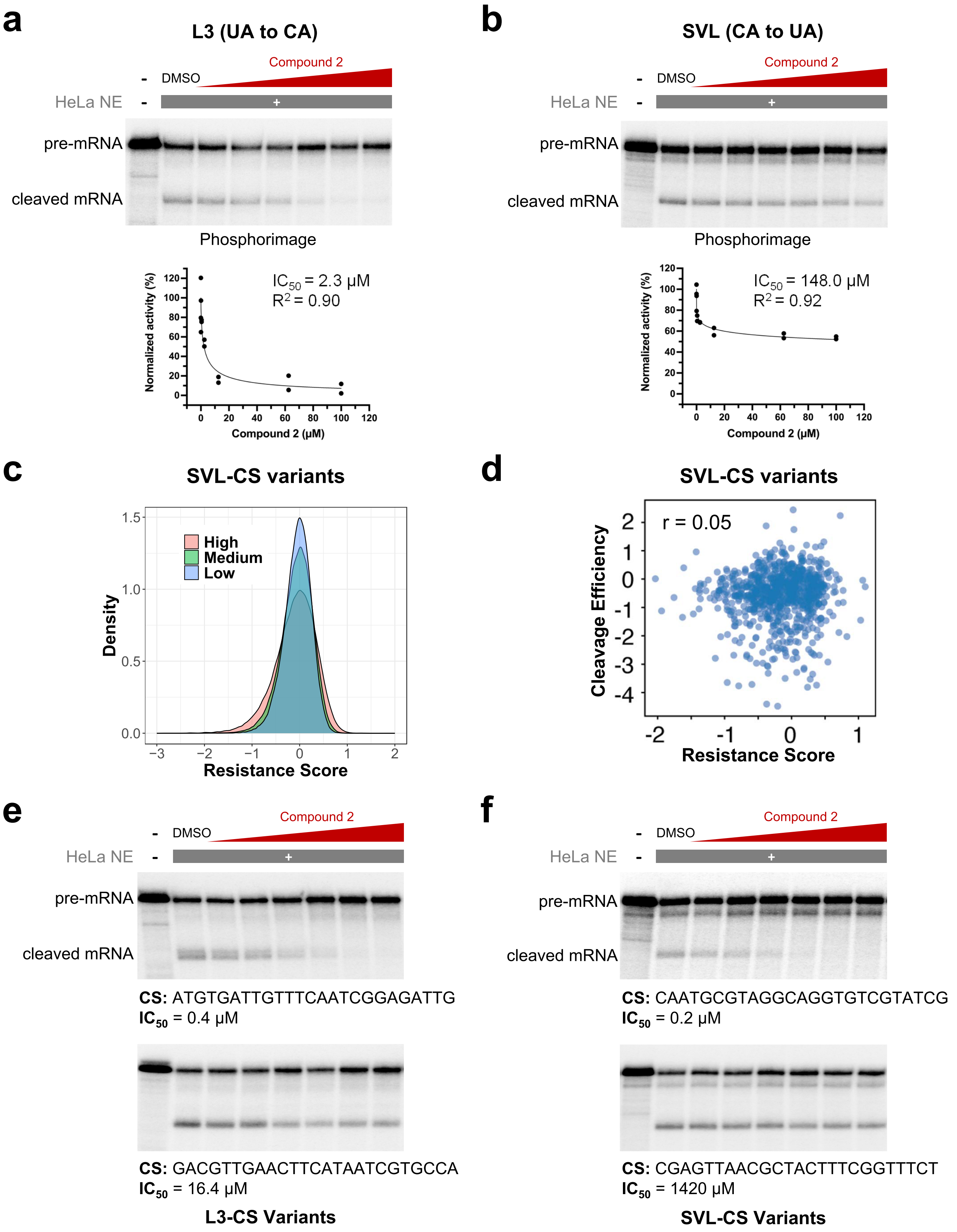
Peer review information *Nature Structural & Molecular Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available. Primary Handling Editor: Sara Osman, in collaboration with the *Nature Structural & Molecular Biology* team.



Extended Data Fig. 1 | Compound 2 does not affect 3′ processing complex assembly on resistant RNA. Electrophoretic mobility shift assay (EMSA) with SVL PAS in the presence of increasing concentration of Compound 2. Same concentrations as Fig. 1a and b were used. n=2 biological replicates.

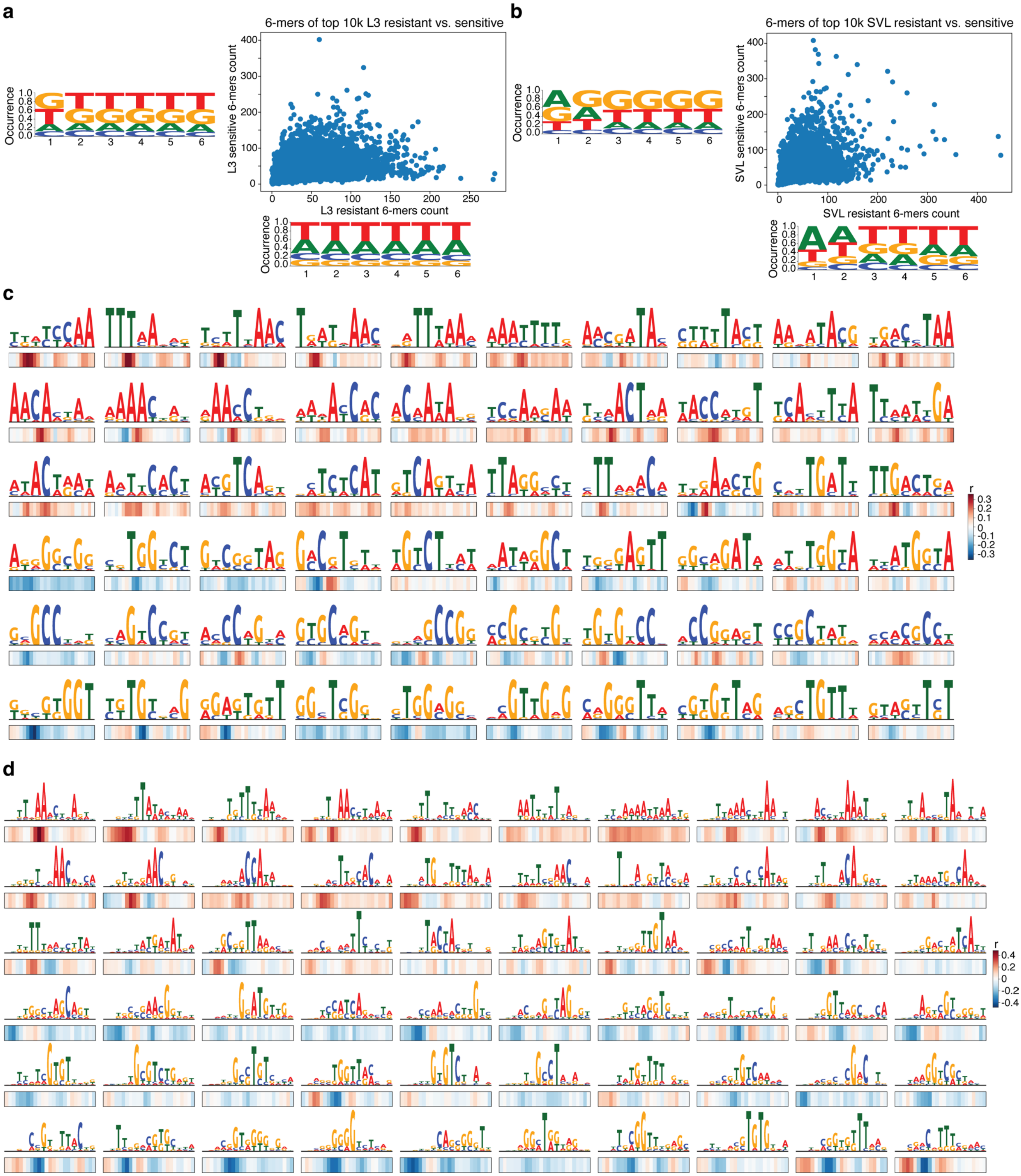


Extended Data Fig. 2 | In vitro cleavage for L3 and SVL chimeras. In vitro cleavage of L3-SVL chimeras 1 (**a**), 2 (**b**), 3 (**c**) and 4 (**d**) with increasing concentration of Compound 2 and their IC_{50}, similar to Figs. 1 and 2. n = 2 biological replicates and both measurements are shown as dots.



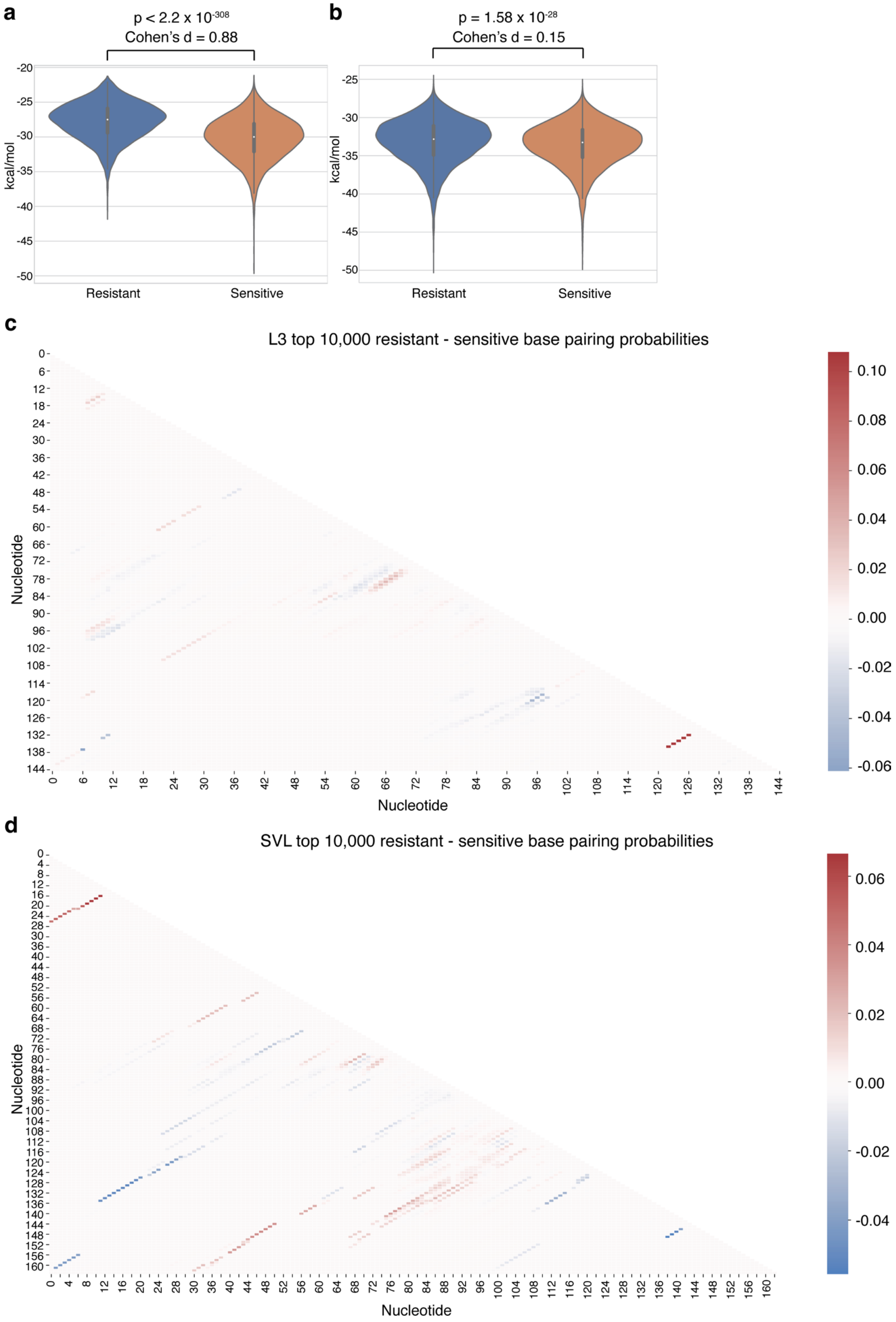
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Development and validation of MPIVA. (a-b) In vitro cleavage on **(a)** L3 (UA to CA mutant) and **(b)** SVL (CA to UA mutant) and their IC_{50}. Compound 2 concentration used is the same as Figs. 1-2. **(c)** A density plot for the resistance scores of all variants in SVL-N23 library. The low, medium, and high groups represent the screens in the presence of 0.5, 2.5, and 12.5 μM Compound 2. **(d)** A scatter plot comparing the cleavage efficiency log(frequency in Library 2/frequency in Library 1) and the resistance score (log(frequency in Library 5/frequency in Library 2)) of SVL-CS variants. Pearson correlation is shown. **(e-f)** In vitro cleavage validation experiment of 4 more RNA (2 sensitive and 2 resistant) from both **(e)** L3-N23 and **(f)** SVL-N23 libraries. The CS region sequence and their IC_{50} is shown. **(a, b)** n=2 biological replicates and both measurements are shown as dots.

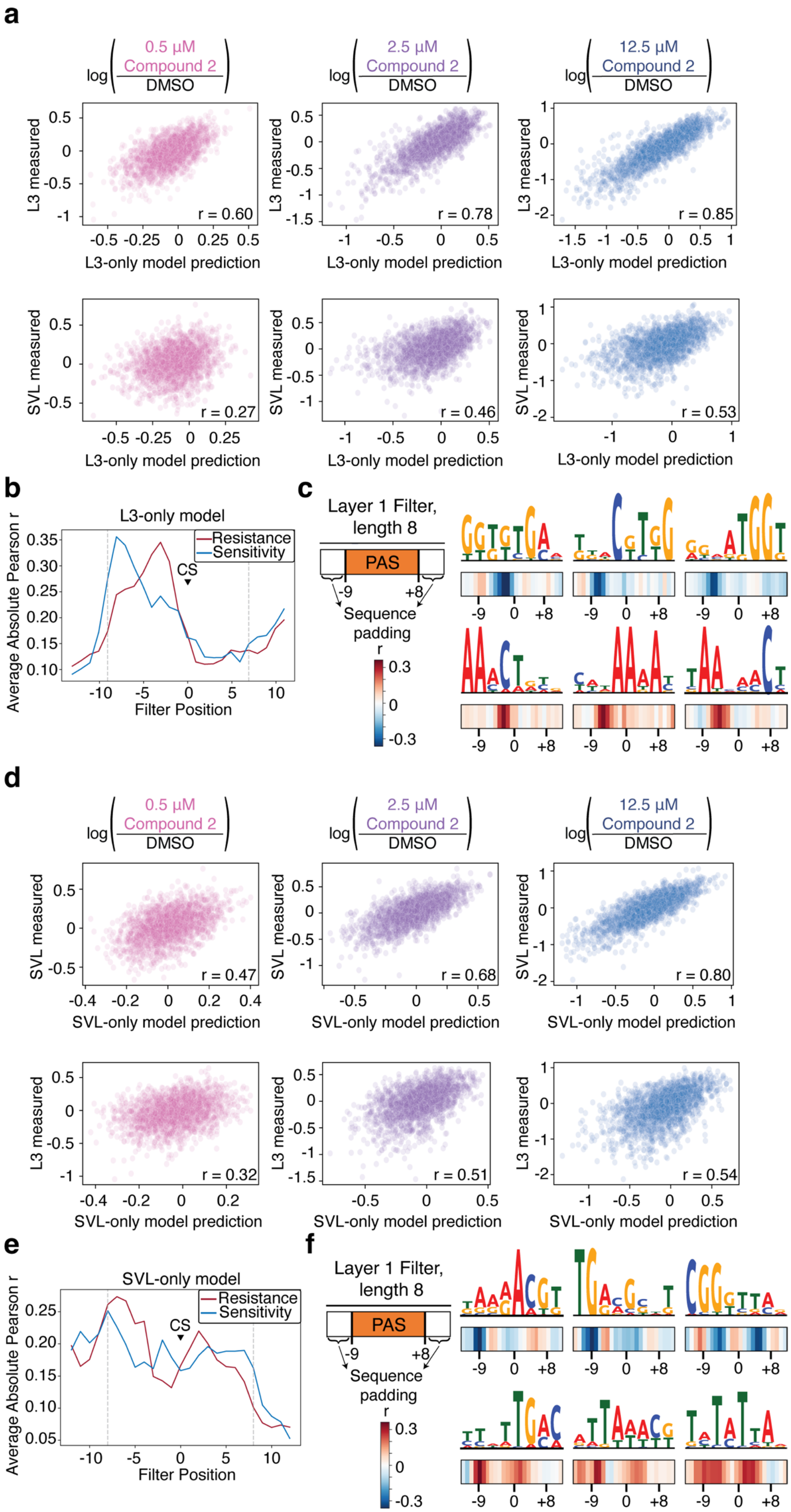


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | 6-mer motif analyses and C3PO learned sequence features from MPIVA. Counts of 6-mers from **(a)** L3 and **(b)** SVL backbones are plotted alongside the nucleotide content of significantly enriched 6-mers in the top sensitive (left logo) and resistant (bottom logo) 10,000 CS variants. Sequence logos use DNA-encoding of RNA nucleotides. Top 10,000 resistant and sensitive sequences were converted into their 6-mer counts. 6-mers in the top 10,000 resistant and sensitive sequences were found to be significant by a binomial test with a null hypothesis of probability of success = 0.25^6 and alternative hypothesis of > 0.25^6. p-value threshold was adjusted by the number of possible k-mers, 4^6, and thus significant 6-mers must have p-values ≤ $0.05/4^6$. The nucleotide content of significant resistant and sensitive 6-mers are shown next to their respective axes. **(c)** C3PO's layer 1 filters' max activation sequence consensus and correlations with 12.5 µM Compound 2 sensitivity predictions. Related to Fig. 4d. Convolutional layers 1 and 2 were analyzed similarly to a previously published analysis of a CNN that predicts alternative polyadenylation (APARENT). In brief, every filter in both convolutional layers were correlated with predictions of drug sensitivity at the 12.5 µM dose. The top 5,000 input sequences from the training set that achieved maximal filter activation were put into a position weight matrix and used to generate position-aware consensus sequence logos. Pearson's r plots of each filter's activations with predicted 12.5 µM Compound 2 sensitivity at each position are plotted below these filter-specific sequence logos. Layer 1 filters are 8 positions wide, and layer 2 filters are 15 positions wide. Note that the convolutional layers in C3PO contain even zero-padding to maintain an input/output size of 25. The padding should be accounted for when analyzing the filters' Pearson r plots. For example in layer 1, the RNA sequences are padded with 4 0's on both the left and right, and the first position in the correlation plots corresponds to 3 0's and 5 nts of the randomized region. **(d)** C3PO's layer 2 filters' max activation sequence consensus and correlations with 12.5 µM Compound 2 sensitivity predictions. Related to Fig. 4e.

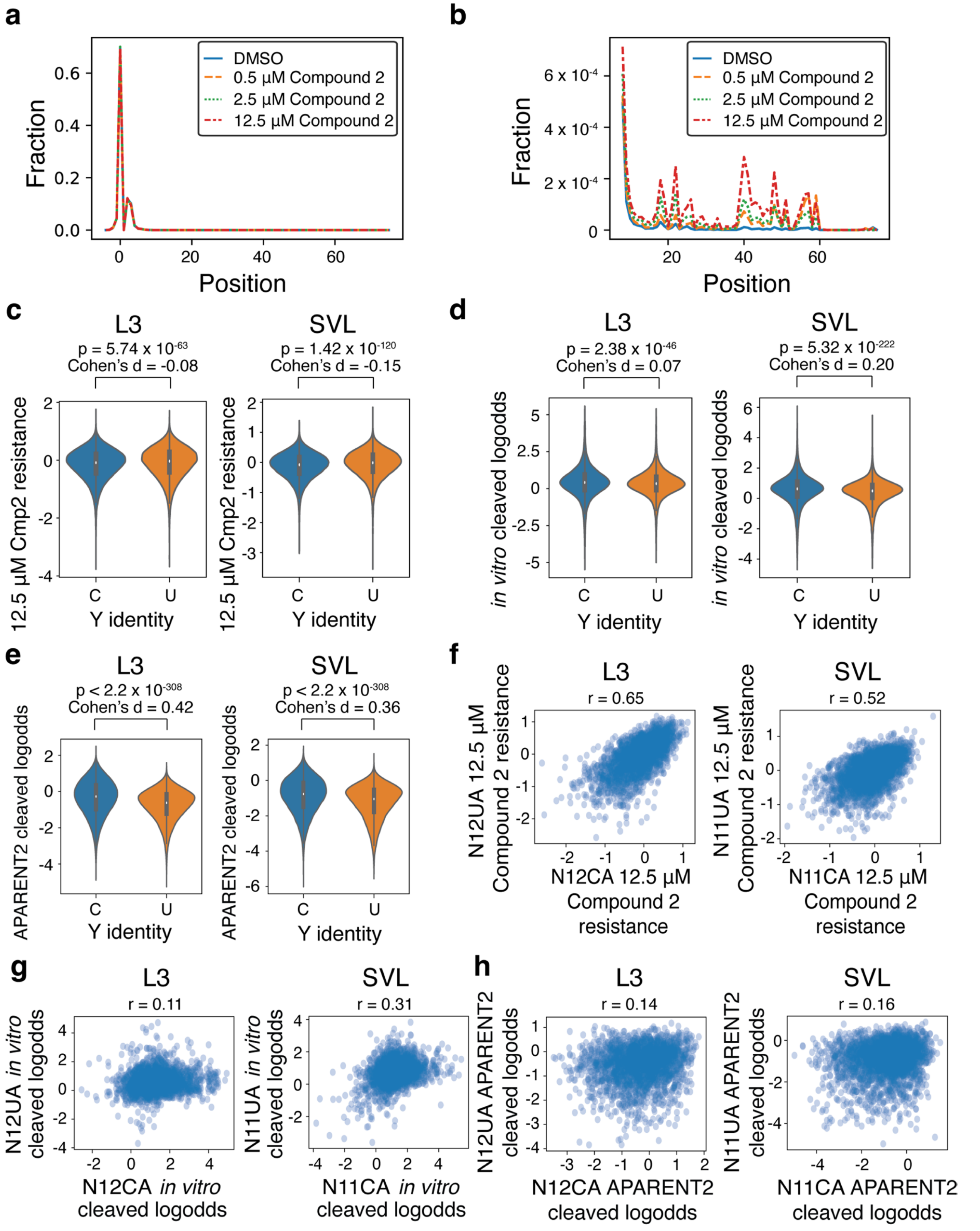


Extended Data Fig. 5 | ΔG of minimum free energy structures and base pairing probabilities of the top 10,000 resistant and sensitive sequences. Comparison of minimum free energy (MFE) structures' of ΔG's from the top 10,000 resistant and sensitive **(a)** L3 and **(b)** SVL sequences. For all boxplots, hinges were drawn from the 25th to 75th percentiles, with the middle line denoting the median, and whiskers with maximum 1.5 interquartile range. The ΔG's are significant with a p-value of $< 2.2 \times 10^{-308}$ for L3 and 1.58×10^{-28} for SVL (two-sided t-test with unequal variance). **(c)** Heatmap of the difference between top 10,000 resistant and sensitive L3 sequences' average base pairing probabilities. **(d)** Same as in panel c but for SVL.

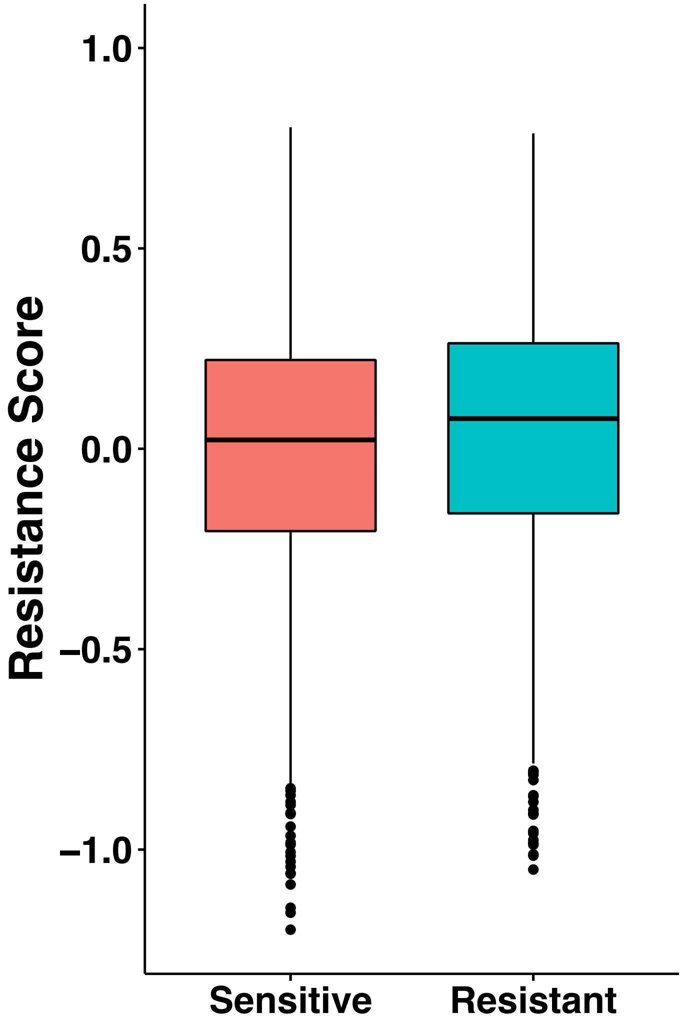


Extended Data Fig. 6 | See next page for caption.

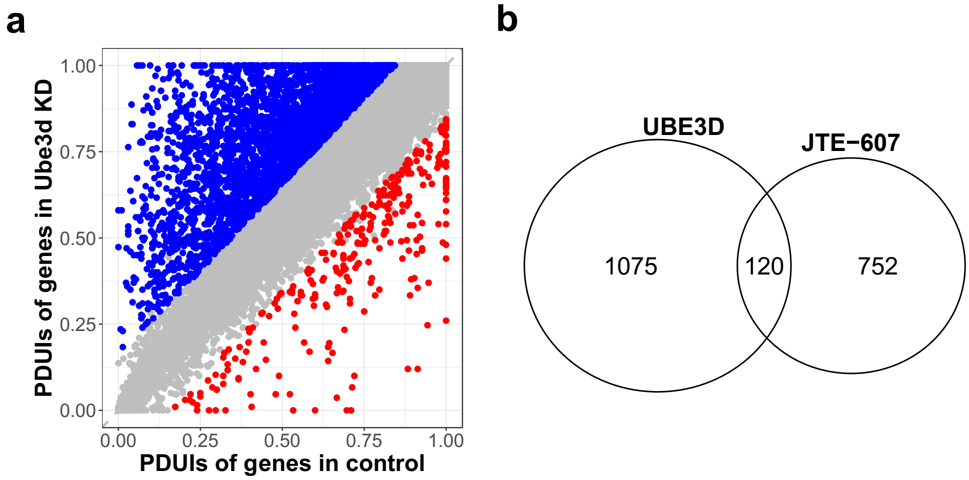
Extended Data Fig. 6 | Performance and interpretation of machine learning models trained exclusively on one MPIVA RNA sequence context. **(a)** Scatter plots of L3-only model performance on predicting drug sensitivity at 3 Compound 2 doses on L3 test sequences (upper) and SVL test sequences (lower). Test sequences include equal number of sequences derived from both the L3 and SVL RNA contexts. **(b)** Plot of average of all L3-only model's layer 1 filters' absolute value of Pearson correlation with 12.5 µM Compound 2 predictions across all positions. These are split into Pearson correlation values associated with resistant, negative, and all 12.5 µM Compound 2 predictions. Dashed gray lines indicate positions at the edge of sequence padding. **(c)** L3-only model's convolutional layer 1 max filter activations with the highest Pearson correlation with 12.5 µM Compound 2 predictions. Sequence logos are plotted on top of per-position absolute value of Pearson correlations with 12.5 µM Compound 2 sensitivity predictions. Filters' Pearson correlations that begin at the canonical cut site in the SVL context are marked, and note that preceding filters may overlap with the designed canonical cut sites. **(d)** Same analyses as in panel a, but for the SVL-only model. **(e)** Same analyses as in panel b, but for the SVL-only model. **(f)** Same analyses as in panel c, but for the SVL-only model.



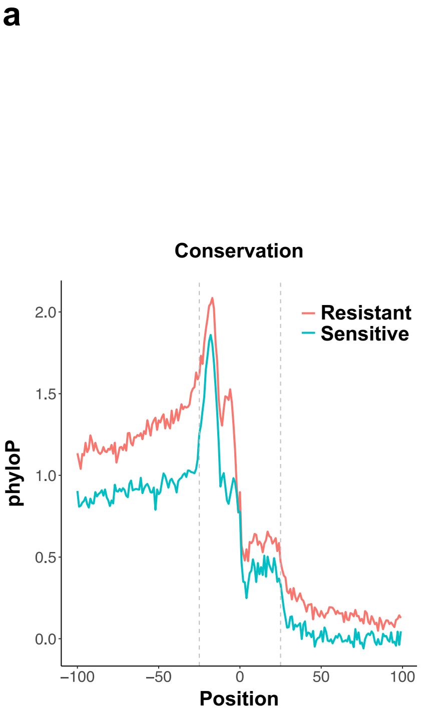
Extended Data Fig. 7 | Cleavage position in MPIVA and effects of YA identity. **(a)** Fraction of cleavage position usage across all 4 MPIVA datasets. Position 0 demarcates expected cleavage position. **(b)** Same as plotted in panel a, but only showing positions +8 and greater. Comparison of Y identity on **(c)** 12.5 µM Compound 2 resistance in the MPIVA datasets (p-value L3, SVL = 5.74×10^{-63}, 1.42×10^{-120}), **(d)** *in vitro* cleaved logodds (p-value L3, SVL = 2.38×10^{-46}, 5.32×10^{-222}), and **(e)** APARENT2-predicted cleaved logodds in the MPIVA datasets (p-value L3, SVL = $< 2.2 \times 10^{-308}$, $< 2.2 \times 10^{-308}$). Two-sided t-tests with unequal variance were used for all statistical tests. For all boxplots, hinges were drawn from the 25th to 75th percentiles, with the middle line denoting the median, and whiskers with maximum 1.5 interquartile range. Scatter plots of **(f)** 12.5 µM Compound 2 resistance, **(g)** *in vitro* cleaved logodds, and **(h)** APARENT2-predicted cleaved logodds of pairs of sequences that share the same sequences upstream of the YA dinucleotide in the MPIVA datasets.



Extended Data Fig. 8 | Comparison of top 1000 sensitive and resistant PASs. 1,000 genes with the most (sensitive) and least significant (resistant) readthrough after JTE-607 treatment were identified based on our 4sU-seq data and the resistance scores of their PAS were predicted by C3PO and compared. Two-sided t-test: p = 0.0063.



Extended Data Fig. 9 | Comparison of JTE-607 treatment with Ube3d knockdown. (a) Poly(A) site usage index (PDUI) in control and Ube3d knockdown cells. Blue dot: genes with significant 3′ UTR lengthening; Red dots: genes with significant 3′ UTR shortening. **(b)** A Venn diagram comparing the genes with significant APA changes in Ube3d knockdown and JTE-607-treated cells.



b

Top 1000 Sensitive PAS Biological Process	Padj
Establishment of protein localization to mitochondria	1.40e-5
Ribosome biogenesis	2.94e-5
Protein polyubiquitination	1.29e-2
Golgi vesicle transport	1.49e-2
rRNA metabolic process	1.74e-2
Histone acetylation	3.32e-2

Top 1000 Resistant PAS Biological Process	Padj
RNA localization	4.40e-7
Ribosome biogenesis	6.00e-7
Regulation of binding	1.04e-5
DNA conformation change	1.31e-3
Cellular respiration	2.21e-3
Intracellular protein transmembrane transport	8.16e-3

Extended Data Fig. 10 | Conservation and gene ontology analyses for JTE-607-sensitive and –resistant PASs. (a) The phyloP sequence conservation score for both resistant and sensitive PASs across different species was calculated and plotted against nucleotide position of the CS. Position 0 is the YA (Y is U or C) cleavage position. **(b)** Gene ontology analyses of genes that contain the top 1000 sensitive or resistant PAS in HepG2 cells. This analysis was done with gProfiler and the top 6 categories that contain between 5 and 500 genes are listed.

nature portfolio

Corresponding author(s): Yongsheng Shi and Georg Seelig

Last updated by author(s): Sep 13, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Cytiva ImageQuant TL v8.2.0.0 (for phosphorimaging), Nanodrop 2000 v1.3.1 (DNA/RNA extraction)
Data analysis	STAR (2.7.3a), Deeptools (3.5.0), BEDTools2 (2.30.0), cutadapt (4.0), RStudio (2022.02.3), edgeR (3.341.1), bbmerge (38), starcode (1.4), RNAStructure (6.4), pingouin (0.5.3), C3PO (https:// github.com/angelamyu/C3PO), APARENT2 (https://github.com/johli/aparent-resnet)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All high throughput sequencing data described in this manuscript, including PAS-seq, 4sU-seq, and massively parallel in vitro assay (MPIVA) data, has been deposited to the GEO database and the accession number is: GSE218977

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	n/a
Reporting on race, ethnicity, or other socially relevant groupings	n/a
Population characteristics	n/a
Recruitment	n/a
Ethics oversight	n/a

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences　　□ Behavioural & social sciences　　□ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were determined according to conventions for genomic analyses (at least 2 biological replicates per condition). Significance level was set to 0.05.
Data exclusions	No data was exlcluded
Replication	Experiments were independently repeated, the number of replicates are presented in the figure legends and/or methods.
Randomization	Randomization was performed in machine learning analysis and the details are presented in Methods. For other analyses, samples were not randomized and were assigned to control and experimental groups based on the experimental treatments.
Blinding	Blinding was not used during experiments or data analysis, because proper controls were used.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
□	☒	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
☒	□	Clinical data
☒	□	Dual use research of concern
☒	□	Plants

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	HepG2: ATCC HB-8065; HeLa S3: a kind gift from Dr. Bert Semler at UC Irvine, original source is ATCC CCL-2.2
Authentication	None of the cell lines were authenticated in the laboratory.
Mycoplasma contamination	All cell lines were tested negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.