Article

Predicting the heat release variability of Li-ion cells under thermal runaway with few or no calorimetry data

Karina Masalkovaitė[1,2], Paul Gasper [2] ✉ & Donal P. Finegan [2] ✉

Accurate measurement of the variability of thermal runaway behavior of lithium-ion cells is critical for designing safe battery systems. However, experimentally determining such variability is challenging, expensive, and time-consuming. Here, we utilize a transfer learning approach to accurately estimate the variability of heat output during thermal runaway using only ejected mass measurements and cell metadata, leveraging 139 calorimetry measurements on commercial lithium-ion cells available from the open-access Battery Failure Databank. We show that the distribution of heat output, including outliers, can be predicted accurately and with high confidence for new cell types using just 0 to 5 calorimetry measurements by leveraging behaviors learned from the Battery Failure Databank. Fractional heat ejection from the positive vent, cell body, and negative vent are also accurately predicted. We demonstrate that by using low cost and fast measurements, we can predict the variability in thermal behaviors of cells, thus accelerating critical safety characterization efforts.

Lithium ion (Li-ion) batteries have helped make many modern inventions practical, from electric vehicles, portable electronics, to reliably powered spacesuits[1]. It is vital that Li-ion batteries and the devices that use them are safe. However, there is always some risk that a cell will undergo thermal runaway (TR) due to challenging operating or environmental conditions, or defects that cause short circuits[2]. Additionally, if one cell in a pack undergoes TR, there is also the risk of TR propagation to neighboring cells that can result in disastrous outcomes[1,3]. Propagation occurs when heat generated by one cell undergoing TR heats neighboring cells, causing TR to spread. Thus it is critical to understand the heat output of Li-ion cells during TR to enable design of safe battery energy storage systems.

The heat output by a Li-ion cell greatly varies between tests and can be influenced by factors such as cell properties, cycle histories, and abuse test conditions[1,2,4–6]. However, even when identical cells are evaluated under identical abuse test conditions, considerable variability in the thermal runaway behaviors is still observed[7,8]. When considering cylindrical Li-ion cells, the heat ejected from the positive end, the negative end, and the cell body are each important to quantify to

design safe, thermal runaway propagation resistant battery systems[9–11]. The fractional breakdown of heat ejected and not ejected (remains in cell casing) is referred to as the fractional heat output, i.e., the fraction of the total heat output that is attributed to ejected material from either end of the cell or emitted from the cell casing itself. These measurements can be recorded using a fractional thermal runaway calorimeter (FTRC), but this equipment is not widely available, and each experiment is time-consuming[5]. Additionally, heat output is variable such that two identical cells undergoing TR in identical environments will have varying heat outputs and behaviors depending on what happens inside the cell, resulting in a distribution of heat outputs from a cell type and possibly outliers that may present rare but hazardous failure modes[4]. To account for this variance, numerous abuse tests and thermal runaway measurements are necessary to understand the range of failure scenarios and thermal behaviors that occur, which is an expensive and time-consuming process[12]. The expense and challenge of characterizing the distribution of behaviors has been highlighted by researchers at Volkswagen and Ford[12,13], and considerably delays adoption of new cell types for use in electric

¹Department of Materials Science and Engineering, Stanford University, Stanford, CA 94305, USA. ²National Renewable Energy Laboratory (NREL), 15014 Denver W Pkwy, Golden, CO 80401, USA. ✉e-mail: Paul.Gasper@nrel.gov; Donal.Finegan@nrel.gov

vehicles. Therefore, a method to predict the distribution of heat output during TR that is time- and cost-effective would be tremendously useful to accelerate safety evaluations of new cell types.

While there are computational models simulating TR[14], physics-based models require detailed knowledge of the cell chemistry and other cell properties that may be proprietary. They are also deterministic and output a single result rather than conveying the real-world complex distribution of occurrences. A data driven approach such as machine learning (ML) may be used to predict the stochastic thermal response of cells but requires empirical data to train. ML models have been commonly utilized in the battery field to shortcut the need for a physical model when making complex predictions, such as predicting internal short circuits[15], state of charge monitoring[16,17], health diagnosis[18,19], future health prediction[20–22]. Several works have also utilized ML to improve battery design[23,24]. However, ML methods need plentiful, high quality, and robust data for training. Such data on thermal behaviors of Li-ion cells during thermal runaway has not been openly available until the Battery Failure Databank[25] was released by the National Renewable Energy Laboratory (NREL) and National Aeronautics and Space Administration (NASA), which presents data from hundreds of FTRC tests providing information on total heat output, the fraction of heat ejected from cells, the mass ejected from cells, all of which will be used in this work. The Battery Failure Databank also hosts many high-speed synchrotron radiography videos of the thermal runaway processes which will not be used in this work. The Battery Failure Databank[25] is the largest public database containing information about batteries undergoing TR; it contains test results on batteries from various manufacturers and twenty-two battery types, totaling to over 350 trials as of November 2023, and continues to expand.

Here, we focus on the use of experimental data that are simple to measure such as mass ejection, and show that these data can be used to predict the complex thermal behaviors that are measured using sophisticated calorimetry techniques like FTRC. Measuring the ejected mass of cells is simple; complex equipment is not required and the throughput of testing can be high. We develop an ML model for predicting the variable fractional heat output of cells undergoing TR using only ejected mass data and the specifications of the cell provided by the manufacturer, avoiding the need for any detailed physical or electrochemical properties, or sophisticated calorimetry techniques. The predictive models were developed using a subset of data found in the Battery Failure Databank[26], which for the first time facilitated a robust experimental dataset for training ML models on thermal runaway behavior. The performance of ML models is studied by cross-validation, training on data from all cell types but one, and testing on data from unseen cell types. Zero-shot predictions are made assuming that no heat output measurements have been recorded, only ejected mass measurements, and allow for qualitative estimation of the heat output from any given cell type without any calorimetry data. One-, two-, …, i-shot predictions, where i FTRC measurements have been performed, enable quantitatively accurate predictions of the mean and variance of both total and fractional heat output with just 0-5 FTRC measurements for most cell types, overcoming previous limitations of models and simulations in predicting single outcomes without capturing the real-world distribution of behaviors. The ability to predict the distribution of the heat output for any given cell type from multiple ejected mass tests and only a small number of heat output measurements is expected to empower energy storage system designers to accurately and rapidly estimate the safety risks of new cylindrical cell Li-ion batteries with minimal expense and effort.

Results
Training data and insights from experiments
Data collection occurred over multiple years and is reported in the Battery Failure Databank from NASA and NREL[25,26], provided in the supplementary data. General trends from the data set have been previously reported[1,2,4]. Since absolute values of total heat output are related to charge capacity, mass, and cell size, analysis of heat output and ejected mass is simplified by normalizing heat output by charge capacity and ejected mass by total mass[5]. Figure 1 shows the distributions of ejected mass fraction, normalized total heat output, and fractional normalized heat outputs (positive, cell body, and negative terminal heat outputs) during TR for each of 8 cell types used for model training in this work. Only commercially-produced 18650 or 21700 format cells tested at 100% state-of-charge with more than 10 samples were considered in this work. Additional data from cells at 100% SOC but with fewer than 10 measurements, at lower SOCs, or after modifying the cell are used as a secondary test set for testing model performance. For both ejected mass fraction and total heat output, the median and variance of values across the cell types varies quite a bit, with a general trend of low ejected mass correlating to low total heat output, however, large variance of the ejected mass fraction does not necessarily correspond to large variance of the heat output (KULR 18650-K330). Of the fractional heat output measurements, the cell body heat output is the most consistent across the cell types, while both the mean and variance of the positive and negative heat outputs varies widely. Note that two cell types (LG 18650-MJ1 and Sanyo 18650-A) were observed to have zero or near-zero negative heat output, while other cell types (KULR 18650-K330) have up to 50% of their total heat output measured on the negative side.

As noted, there is a general correlation observed between the total ejected mass fraction and total normalized heat output, prompting this investigation to study whether heat output may be predicted simply through measurements of ejected mass, given the existing FTRC data reported in the Battery Failure Databank. Figure 2a shows a linear regression of total normalized heat output from ejected mass fraction, indicating substantial noise but an overall strong correlation. However, when observing the trend for each cell type independently, as in Fig. 2b, the data set as a whole has a stronger relationship between ejected mass and heat output than for any cell independently, and substantial variance in the strength of the correlation begins to appear. For example, LG 21700-M50 cells have a positive slope and relative narrow confidence intervals, indicating a strong correlation between ejected mass and heat output, while the LG 18650-MJ1 cells show zero slope with wide confidence intervals, indicating a near zero correlation. This dependency of the FTRC results on the cell type thus requires treating cell types as independent from one another, however, the global trend suggests that it should be feasible to make reasonable heat output predictions for new cell types with a very small number of FTRC measurements.

Predicting thermal behavior of cells
To train a ML model to predict the normalized fractional heat output of batteries undergoing TR, The data shown in Fig. 1 was carried through the processing steps shown in Fig. 3. No filtering to remove outliers was performed, as high heat outputs or high ejected mass fractions are critical to keep in the data set, as they are representative of worst-case failure events. Predicted values were normalized cell body, negative, positive, and total heat output using a chain regression, as the total heat output is a combination of all the fractional heat outputs. The chain regressor was also found to result in lower error than treating each heat output independently. Features used to make predictions included cell metadata, FTRC experiment notes, and mass data after TR, in both absolute and normalized values. For evaluating the performance of models on predicting data from new cell types, data was split into training and test sets where all data from one cell type was held out for testing, and all other cells used for training; predictions for the test cell use zero measured heat output values, i.e., 'zero-shot' models. The impact of conducting additional FTRC measurements on new cell types for calibrating heat output predictions
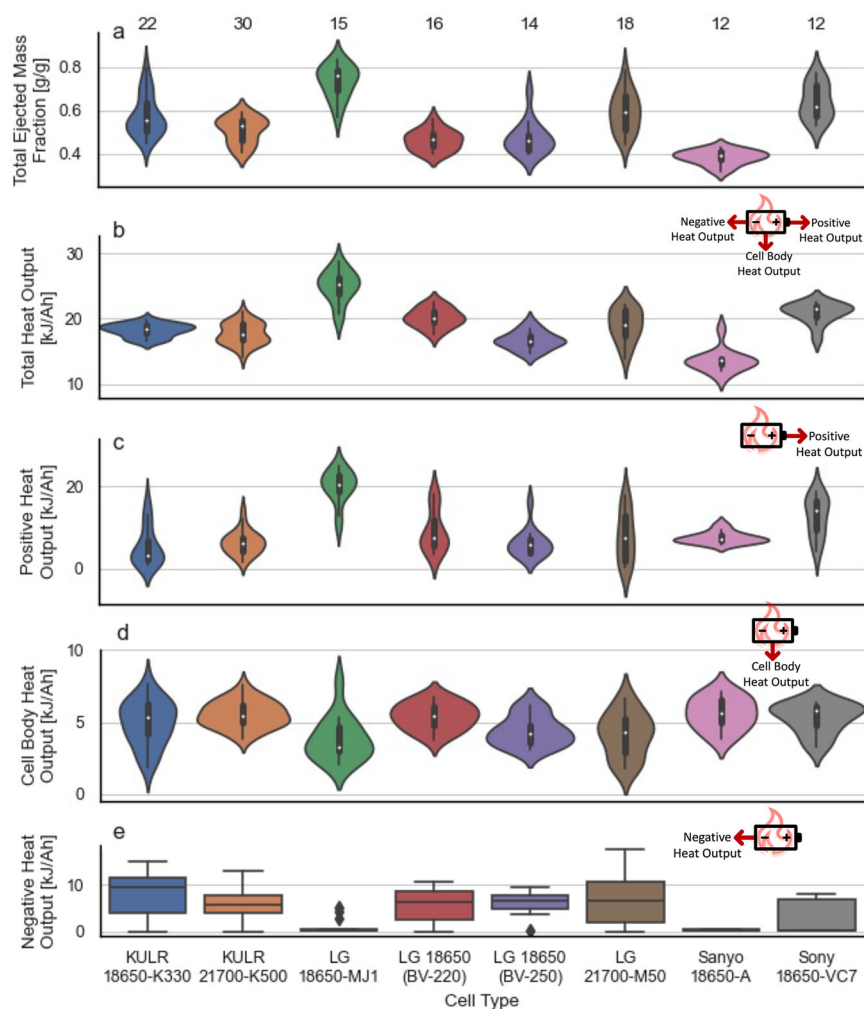
Fig. 1 | Distributions of heat and mass output behaviors of cells. Distributions of (**a**) total ejected mass fraction, (**b**) total normalized heat output, and fractional normalized heat outputs from (**c**) positive, (**d**) cell body, and (**e**) negative cell faces during TR for 8 of the 22 cell types from the Battery Failure Databank[26]. The number above the distributions in (**a**) indicates the number of samples available in the data set. All violin plots have an associated box-and-whisker diagram in their centers.
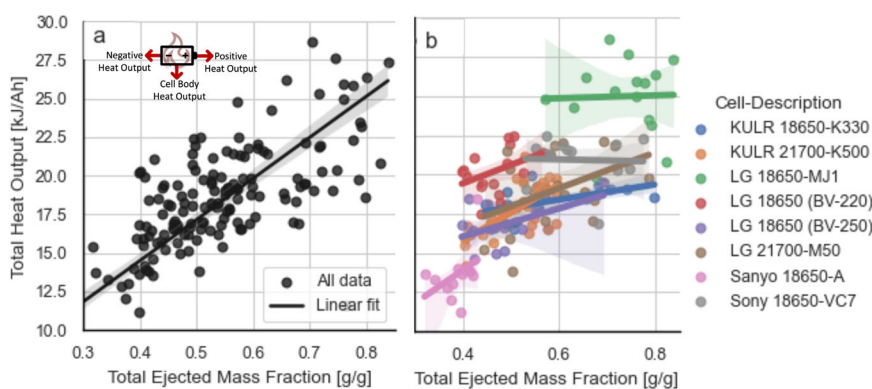

Fig. 2 | Linear regression fits for heat and mass output data. a Linear regression of total normalized heat output from total ejected mass fraction with 95% confidence interval of the entire data set conveyed as a semi-transparent shading. **b** Local linear regressions for each cell type.

was investigated using 'i-shot' models, where i samples from the test cell type were added to the training set. However, this introduces a sampling bias into the model training process. For example, there are 22 possible 1-shot models for the KULR 18650-K330 cell type which can have varying effects on the model, depending on which specific FTRC trial is added to the training set. Three model architectures were considered: a baseline model, a support vector machine (SVM) model,

and a gradient-boosted tree model, XGBoost[27]. The baseline model is a linear regression model using only ejected/retained mass features, using a global regression on all training cells to make zero-shot predictions and a local linear regression on the i samples for each i-shot prediction, exactly as shown in Fig. 2. Comparison of the baseline, SVM, and XGBoost models thus demonstrates the benefit of using prior FTRC measurements as well as cell metadata for making
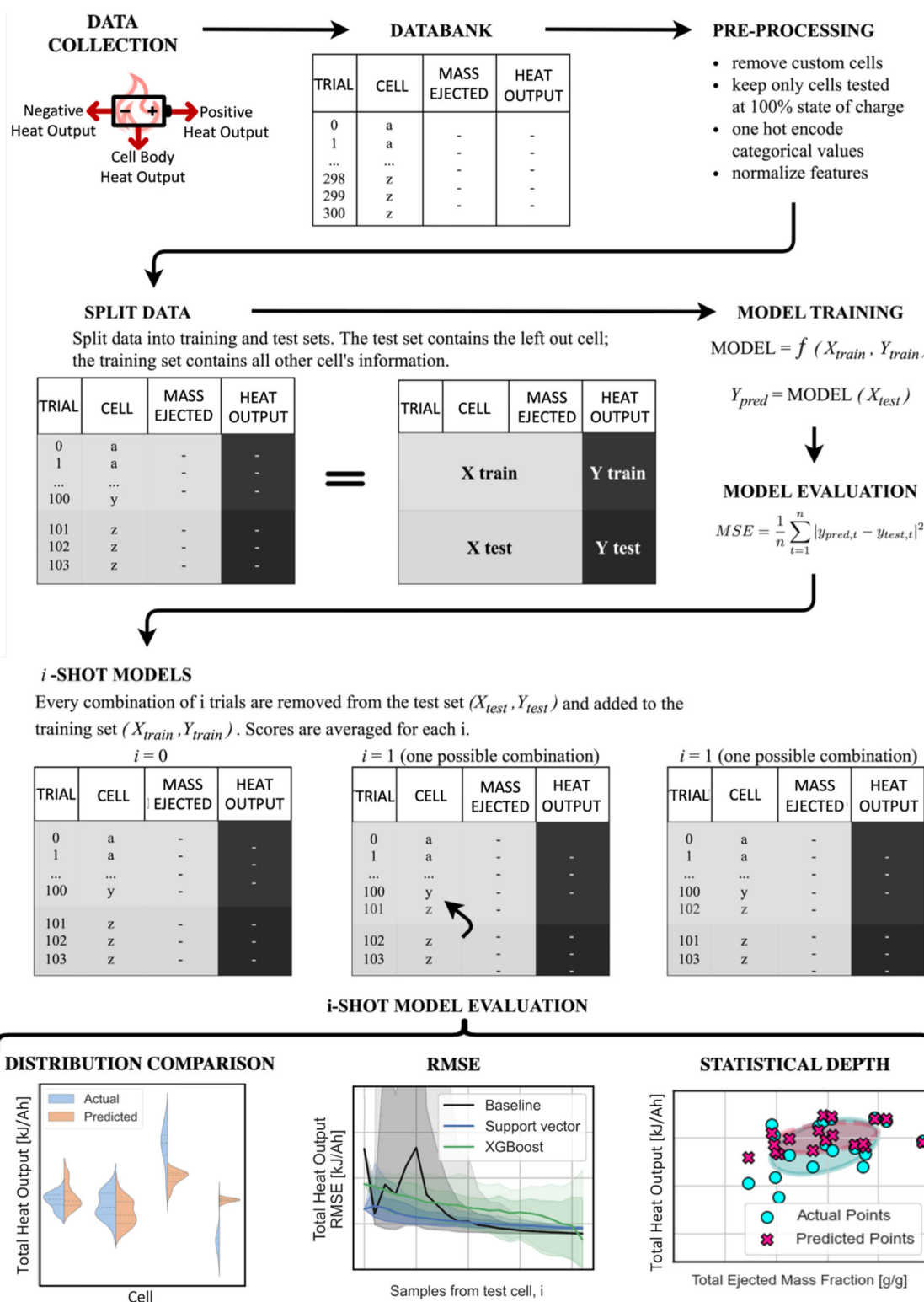
Fig. 3 | Overview of data-treatment process. The data treatment process including pre-processing, model development, and model evaluation.

zero-shot and *i*-shot predictions of heat output during TR events, as opposed to treating each cell type as an independent experiment.

Figure 4a–h compares the RMSE between baseline, SVM, and XGBoost models for each cell. Zero-shot SVM models perform better than baseline for 4 of 8 cell types, comparably to baseline for 2 of 8 cell types, and worse than baseline for only 2 of the 8 cell types. At *i* = 1 to *i* = 5, the benefit of the SVM model becomes clear: with only a few FTRC measurements, the SVM model estimates are near converged with

those made using all of the data. Cells with the worst *i* = 0 predictions show the largest improvements at *i* = 1, demonstrating that heat output predictions for 'outlier' cell types can be calibrated using a single FTRC measurement. SVM confidence intervals rapidly narrow, demonstrating that reasonable heat output predictions can be made for a population of cells while being tolerant to whichever subset of that population is used for measuring fractional heat output in the FTRC. In comparison, the baseline model error is dramatically higher, with large confidence
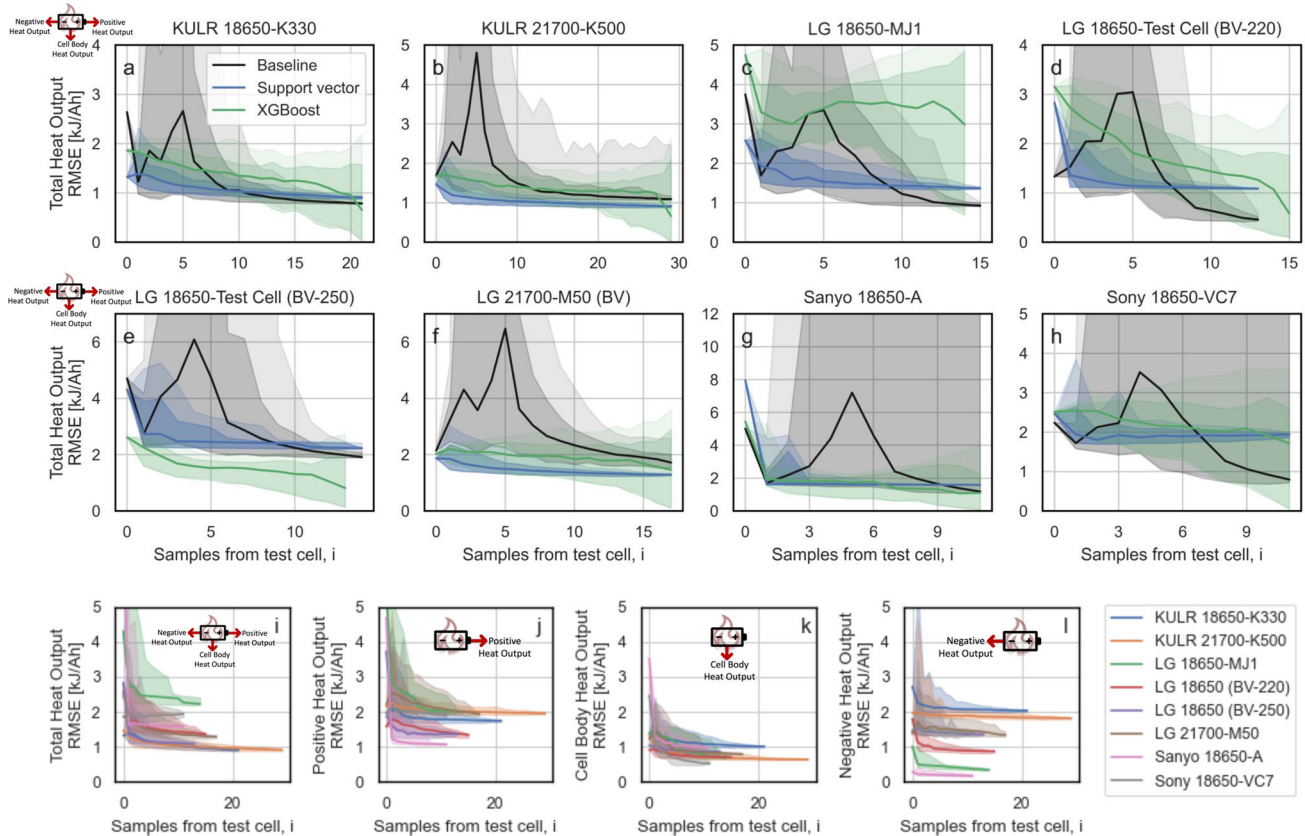
Fig. 4 | Root mean squared errors (RMSE) for predicted heat outputs from cells.
a–h RMSE for the total predicted heat outputs for $i = 0$ to $i = n$ for each cell: (**a**) KULR 18650-K330, (**b**) KULR 21700-K500, (**c**) LG 18650-MJ1, (**d**) LG 18650 (BV-220) (**e**) LG 18650 (BV-250), (**f**) LG 21700-M50, (**g**) Sanyo 18650-A, and (**h**) Sony 18650-VC7. For baseline (black), SVM (blue), and XGBoost (green) models, median RMSE for each i is denoted using a line, while 1σ and 2σ confidence intervals are shown using the dark and light shaded regions. **i–l** RMSE of SVM models for $i = 0$ to $i = n$ for each cell with 2σ confidence intervals: (**i**) normalized total heat output, (**j**) normalized positive heat output, (**k**) normalized cell body heat output, (**l**) normalized negative heat output.

intervals, because treating the TR heat output from each cell type independently means that making any estimate of TR heat output requires a significant number of samples, and the estimated heat outputs for the entire population of that cell type vary wildly depending on which samples are characterized using the FTRC. Performance of the SVM model does vary across cell types, with some cells showing an RMSE of less than $1 kJ/Ah$ at $i = 0$ (KULR 18650-K330), or $\tilde{5}\%$ error compared to the average total normalized heat output for the data set of about $20 kJ/Ah$, while others barely converge below an RMSE of $2 kJ/Ah$ even when trained on all samples (LG 18650-MJ1, Sanyo 18650-A, Sony 18650-VC7). The convergence of baseline and SVM models to similar error at high numbers of training samples suggest that the SVM model is well fit, i.e., not overfit, as an overfit model would begin to report high error on test data as more samples are added to the training data. Comparing the two machine-learning models, SVM and XGBoost, SVM appears to be more accurate and have lower uncertainty across most cell types, and all further results will report predictions from or analysis of SVM models. Similar performance for the SVM model trained on all data is seen on a secondary test set, which is comprised of cells with fewer than 10 measurements or cells that were modified prior to the FTRC measurement, demonstrating that the set of cells used for training extrapolates usefully to new data (Supplementary Fig. 1).

Figure 4i–l compares SVM model RMSEs between normalized total, positive, cell body, and negative heat outputs for each cell type. As noted previously, the worst total heat output RMSE at $i = n$ (LG 18650-MJ1, $2.2 kJ/Ah$) is about twice that of the best cell (KULR 18650-K330, $1 kJ/Ah$). This trend is shared for the positive ($1–2 kJ/Ah$) and cell body ($0.5–1 kJ/Ah$) heat outputs, however, the RMSEs for the negative

heat outputs vary by more than 10 times (about $0.2–2 kJ/Ah$), suggesting that negative heat output does not have a consistent relationship to ejected mass during TR for most cell types. Cell body heat output is predicted with the highest accuracy and with high confidence even at $i = 1$, as was expected from analysis of Fig. 1d because cell body heat output distributions show the least variability across the cell types, that is, it is the easiest heat output to learn accurately from FTRC measurements conducted on other cells.

The predicted versus actual distributions of the normalized total heat output are shown in Fig. 5 for $i = [0, 1, 3, 5]$. As with the RMSE results in Fig. 4, the distributions of half the cell types are accurately predicted at $i = 0$. Of those distributions that are not being well predicted, the zero-shot model predictions show two consistent traits: being biased towards the global mean of the data set, and having much lower predicted variance than the actual variance. At $i = 1$, the impact of the selected sample for training can be clearly seen as the means of the predicted distributions vary substantially, especially for cell types with higher or lower than average total heat outputs. As i increases, approaching the total number of samples n, the means of the predicted distributions converge towards that of the actual distribution, However, for those cells that are 'harder' to predict (LG 18650-MJ1, LG 18650 (BV-250), Sanyo 18650-A), the variability of the heat output shown by the actual distribution is never learned. As noted when describing Fig. 2b, certain cell types have no statistically significant relationship between total ejected mass and total heat output during TR (LG 18650-MJ1), so accurately predicting the actual heat output distribution from ejected mass data is impossible, even with a better performing model.
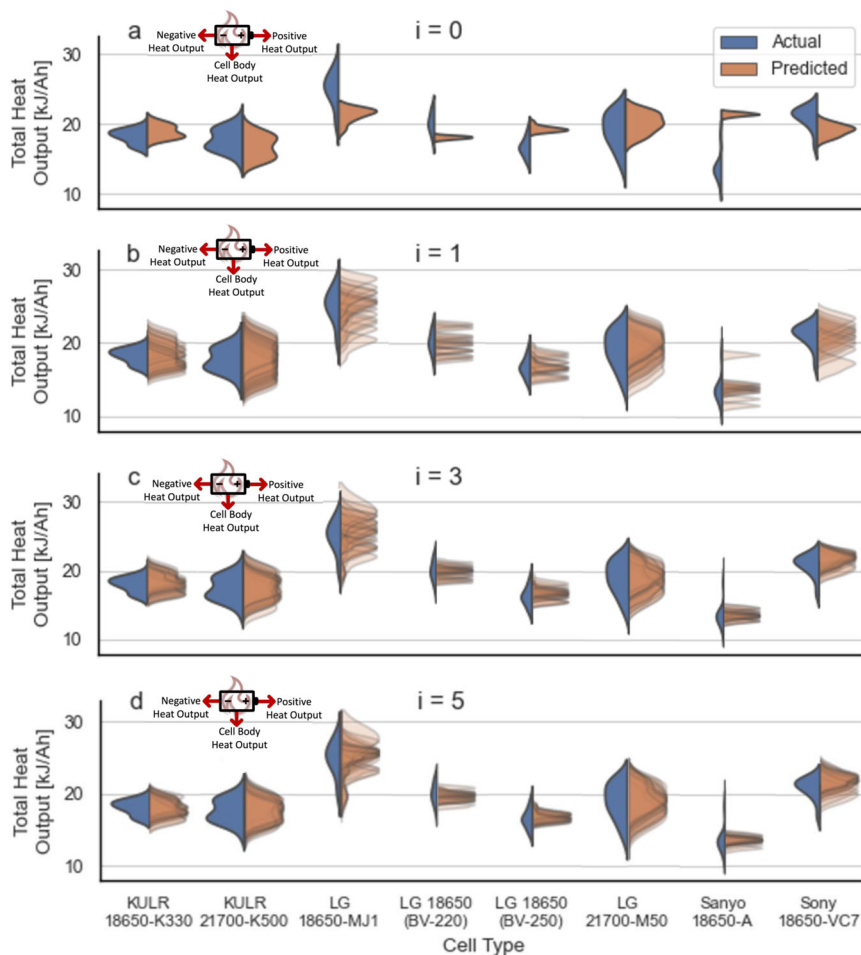
Fig. 5 | Actual and SVM predicted distributions of normalized total heat outputs from cells. Actual and SVM predicted distributions for (**a**) $i = 0$, (**b**) $i = 1$, (**c**) $i = 3$, and (**d**) $i = 5$ models. For $i > 0$, a maximum of 20 random results from the possible $\binom{n}{i}$ training sets are shown for visual clarity.

Predicting heat output from cells sent into thermal runaway at varying SOCs was tested using available data. Both heat and mass output are strongly correlated with SOC, with higher SOCs resulting in more extreme thermal runaway events, as would be expected. While a model trained on only 100% SOC does not extrapolate well to low SOCs, including even a small amount of low SOC data in the model training results in accurate extrapolations to new cell types. See the Supplementary Information and Supplementary Fig. 4 for more details.

Overall, it has been demonstrated that ejected mass values and cell metadata can be used by ML models to accurately predict the average total and fractional heat output. However, while the mean of the distributions was predicted quite well, it is important to also get an initial estimate of outlier scenarios, i.e., the infrequent cells that undergo thermal runaway with anomalously high heat output.

Performance on predicting outliers
Outliers of most interest are the infrequent cells that produce anomalously high heat during thermal runaway. The high heat outliers are important to capture to guide design of battery systems that are resistant to thermal runaway propagation. As found in previous work by the authors, the thermal behavior of cells is strongly influenced by not only how much mass is ejected but when that mass is ejected during the thermal runaway process[4], i.e., how hot the mass is during thermal runaway. This complex dependence can lead to some failure scenarios being considerably more hazardous than others, and predicting the distribution of responses needs to include infrequent outlier behaviors. It has already been observed that some cells have

higher than average error (LG 18650-MJ1, Fig. 4i) or have predicted distributions with obviously lower variance than the actual distribution (Sanyo 18650-A, Fig. 5). Here, we use Kullback-Leibler (KL) divergence to quantify the similarity between the predicted and actual heat output distributions, and utilize Mahalanobis plots to determine some of the sources of error.

Figure 6 shows the KL divergence for the normalized total heat output predictions for each cell. As opposed to RMSE, the KL divergence more clearly shows which cell types have accurately predicted distributions, rather than just accurately predicted means, as the best and worst KL divergence differ by 4 times versus the 2 times of RMSEs, additionally, the worst KL divergence and worst RMSE are from different cells (Sanyo 18650-A and LG 18650-MJ1, respectively), prompting further investigation of both. The KL divergence also shows continual learning for all cells with more data, obviously decreasing in value and confidence interval for most cell types even up to $i = 25$, while the RMSE values plateau at $i = 5$ to $i = 10$. Similar behavior is seen on cells from the secondary test set (Supplementary Fig. 2).

The Mahalanobis plots shown in Fig. 7 demonstrate model behavior on two example cells, one where increasing i values results in continual improvement (LG 21700-M50, Fig. 7a–c), and one where increasing i values shows little change after $i = 1$ (Sanyo 18650-A, Fig. 7d–f). When examining Mahalanobis plots, consider that a narrow ellipsoid corresponds to a strong correlation between the two variables, while a more circular ellipse corresponds to no correlation, i.e., two independent random distributions. For the LG 21700-M50 cell, it is clear that the actual distribution shows a correlation between total

heat output and total ejected mass, though the zero-shot case is not able to perfectly capture this correlation, with both the mean and variance of the heat output predicted poorly. At $i = 3$, the mean heat output is predicted correctly with slightly improved variability, and by $i = 10$ both the mean and variance of the heat output are near correct; as expected, since 10 of the 18 samples shown were used for model training. In comparison, the actual Sanyo 18650-A cell data shows little correlation between total heat output and total ejected mass, so the model can only learn the mean heat output, and has no input features that enable it to accurately predict the variability of the observed heat output even when training on 10 of the 12 samples. The SVM model, which is forced to learn a relationship between ejected mass and heat output, always predicts a distribution with a strong correlation
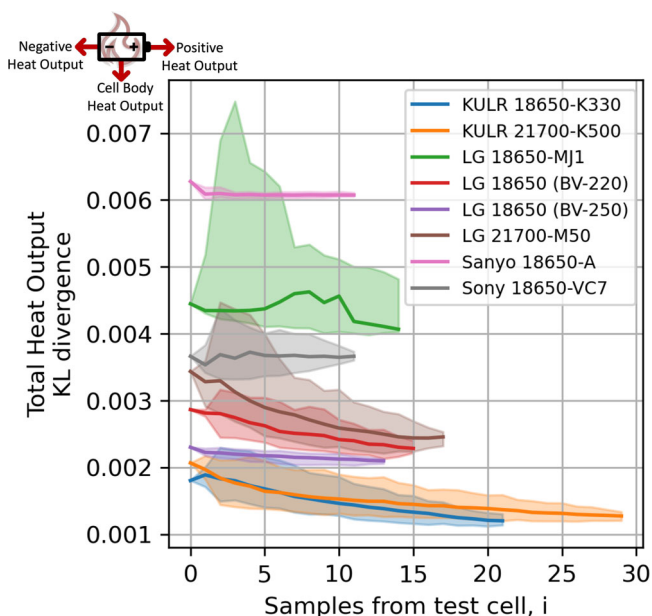
between ejected mass and heat output, especially for cells like the Sanyo 18650-A, which has a small range of ejected mass values relative to the rest of the data set.

Despite some cells showing near zero correlation between total ejected mass and total heat output, the fractional ejected mass and heat output values may still show strong correlations, enabling mass ejected data to be used to predict the variance of the fractional heat outputs accurately, even if the variance of the total heat outputs unlearnable. Figure 8 shows Mahalanobis plots for total and fractional mass/heat for the LG 18650-MJ1 cell at $i = 0$ and $i = 3$. The ellipse for the actual distribution of total ejected mass and total heat output (Fig. 8a,e) is nearly a perfect circle, showing near zero correlation. However, the fractional ejected mass and heat output values are all either partially or strongly correlated, with very narrow ellipses for the cell body and negative fractional values. Thus, as i increases, even as the total heat output prediction does not substantially improve, the fraction heat output predictions are able to learn both the mean and variance of the actual distribution.

The relative impact of each feature on model predictions is quantified using Shapley additive explanations (SHAP)[28]. Figure 9 reports the mean absolute SHAP value for each input feature and each regression model. Cell type has a modest impact on regression outputs, with the exception of the LG 18650-MJ1 type, which has a strong impact on the cell body and total heat output predictions. This cell type is the only cell type in this data set that contains a graphite-silicon composite electrode, which has higher specific energy density than graphite, thus resulting in a higher heat output. None of the cell types have much influence at all on the negative heat output, suggesting that the negative ejected mass fraction and heat output are not strongly related to how manufacturers design the cell casings. Of the cell design details and trigger mechanisms, the presence of a top vent (positive side) and the cell capacity have the largest impacts, with the presence of a top vent most strongly impacting positive heat output and total heat output. The presence of a bottom vent has almost zero impact on model predictions; this may be simply because this feature is redundant with the 'Bottom Vent (BV) actuated' feature, as the bottom vent can only actuate if it isn't there, and if it doesn't actuate, it's as if the bottom vent wasn't actually there. The experimentally measured values, that is, the ejected mass quantities, are the most important features for model prediction. Both absolute (raw) and relative

Fig. 6 | Kullback-Leibler (KL) divergence of SVM model predictions of normalized total heat outputs. KL divergence for $i = 0$ to $i = n$ for each cell with 2σ confidence intervals conveyed as shaded regions.
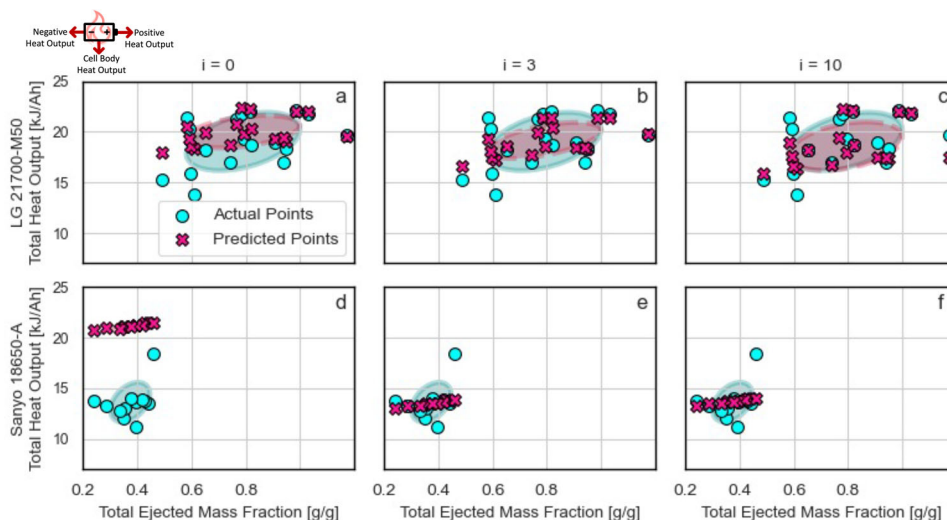

Fig. 7 | Mahalanobis plots of the total ejected mass fractions versus actual and SVM predicted normalized total heat outputs. Mahalanobis plots for the (**a–c**) LG 21700-M50 and (**d–f**) Sanyo 18650-A cells for (**a, d**) $i = 0$, (**b**, e) $i = 3$, and (**c, f**) $i = 10$. Blue shading marks the area of actual points with the highest mahalanobis depth. Red shading denotes the same for the predicted points. Blue and red ellipses, for actual and predicted distributions respectively, cover 66% of the distribution. For $i = 3$ and $i = 10$ plots, only one random set from the $\binom{n}{i}$ possible sample sets is shown for clarity.
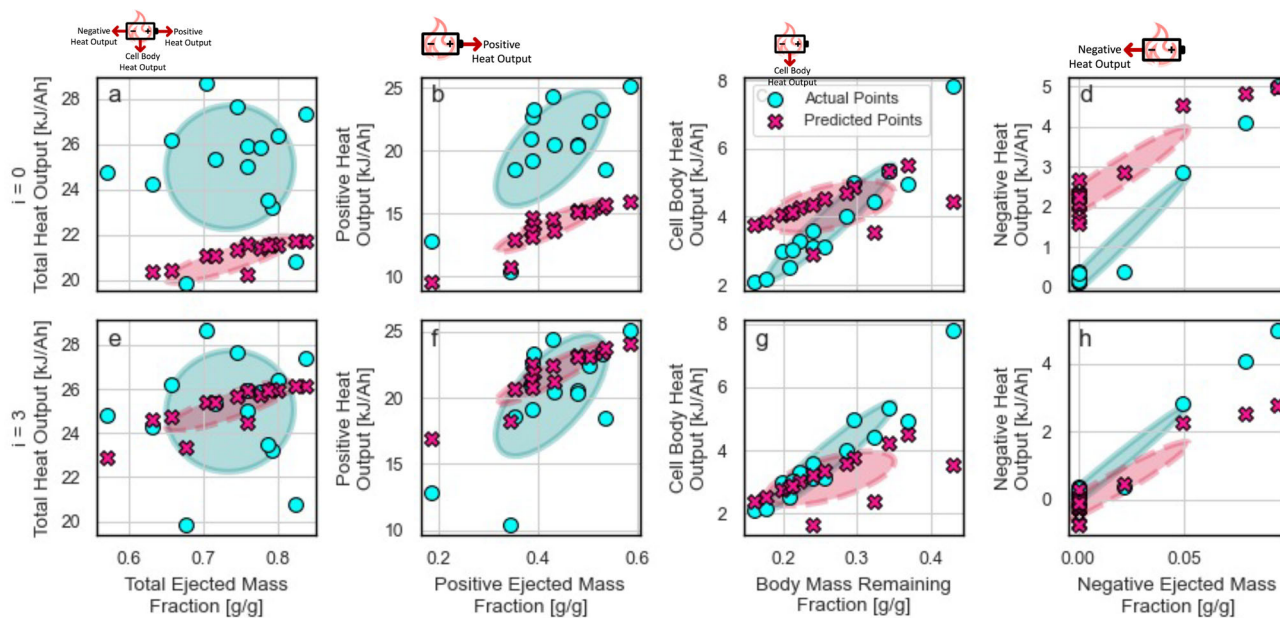
Fig. 8 | Mahalanobis plots for the LG 18650-MJ1 cell of the relative ejected/retained mass features versus the SVM predicted and actual normalized heat outputs. Predicated and actual heat outputs and masses ejected for (**a**–**d**) $i = 0$ and (**e**–**h**) $i = 3$: (**a**, **e**) total, (**b**, **f**) positive, (**c**, **g**) cell body, and (**d**, **h**) negative. For $i = 3$ plots only one random set from the ($\binom{n}{i}$) possible sample sets is shown for clarity. Blue shading marks the area of actual points with the highest mahalanobis depth. Red shading denotes the same for the predicted points.

(calculated) values are important, with the post-test cell-body mass and the negative ejected mass fraction having the highest impacts. The impact of these features on model predictions is intuitive, with the post-test cell-body mass being the most important feature for the total heat output, while negative ejected mass fraction is the most important for the negative heat output. Force plots for each feature and regression model, shown in Supplementary Fig. 3, give more insight into the relationship between features and model outputs. For instance, low post-test cell-body mass is associated with high total heat output, as high amounts of ejected mass is correlated with both low post-test cell-body mass as well as heat generation. Overall, the SHAP analysis demonstrates that the regression models are influenced sensibly by the features, giving confidence that heat output can be predicted using ejected mass and cell details.

Discussion

Like most machine learning approaches on predicting battery behaviors in literature to-date, the technique presented here is limited to the specific set of training data available. The tests used for training mostly contained similar cathodes (varying stoichiometries of $LiNiMnCoO_2$ and $LiNiCoAlO_2$) and similar anodes (graphite with some cells expected to include a small mass fraction of SiOx). Therefore, this training set and model, in its current form, should only be applied to similar cell types and test conditions, for example, see results on the secondary test set cells in Figures S1 and S2. Diverging from these conditions, like to cells with $LiFePO_4$ cathodes or pure Si/SiOx anodes, cells at the end of cycle lives with complex histories, cells at lower states of charge, and cells undergoing different abuse conditions such as overcharge or nail penetration, is not recommended. As the Battery Failure Databank continues to expand over the coming years, additional chemistries, states of charge, and cycle histories may be added, which may give further confidence for predicting behaviors in new realms. Other research groups are encouraged to make their thermal runaway test data open-access, in the hope that collectively, a robust global resource of high-quality data becomes available to strengthen this predictive approach.

In its current form, this approach is anticipated to find many applications. There is an abundance of 18650 and 21700 cell models to choose from on the market, many of which have similar chemistries to the cells evaluated here but with varying power- and energy-densities due to differences in engineering and minor changes in composition. The approach provided here can be used to accurately estimate the distributions of total and fractional heat outputs of such cells by leveraging a large number of cheap and rapid measurements of ejected mass; for example, 5 FTRC measurements could be used to establish the relationship between ejected mass and heat output for a given cell type, and then tens to hundreds of ejected mass measurements could be used to with the machine-learned model to estimate the variability of heat output from thermal runaway events. This can give a guideline for benchmarking cells with regard to their thermal behaviors and therefore quickly assess the suitability of cells for specific applications or for investigating the impact of design changes on cell safety. For example, the secondary test set contains measurements on LG 18650 M36 cells with modified cell casing thickness and with or without a bottom vent, and the SVM model is able to capture the variance and magnitude of heat output from each cell type with low uncertainty with 5 or fewer FTRC measurements. However, while quick and easy predicted estimations are valuable, this approach should not replace conducting actual experiments and measuring the real distributions of behaviors. The results shown here demonstrate that about 5 FTRC measurements are necessary to be confident that the ejected mass data will lead to accurate estimations of the distribution of heat outputs. For cell types where there is no statistical relationship between ejected mass and heat output after 5 FTRC measurements, the most effective method to determine heat output variability is simply to perform as many more FTRC measurements as possible. For example, several of the cells in the secondary test set show no improvement in KL-divergence when adding FTRC measurements to the model training ($i = n$ models), suggesting the heat output from these cells cannot be estimated using ejected mass, and instead would require further FTRC measurements to confidence predict heat output variability.
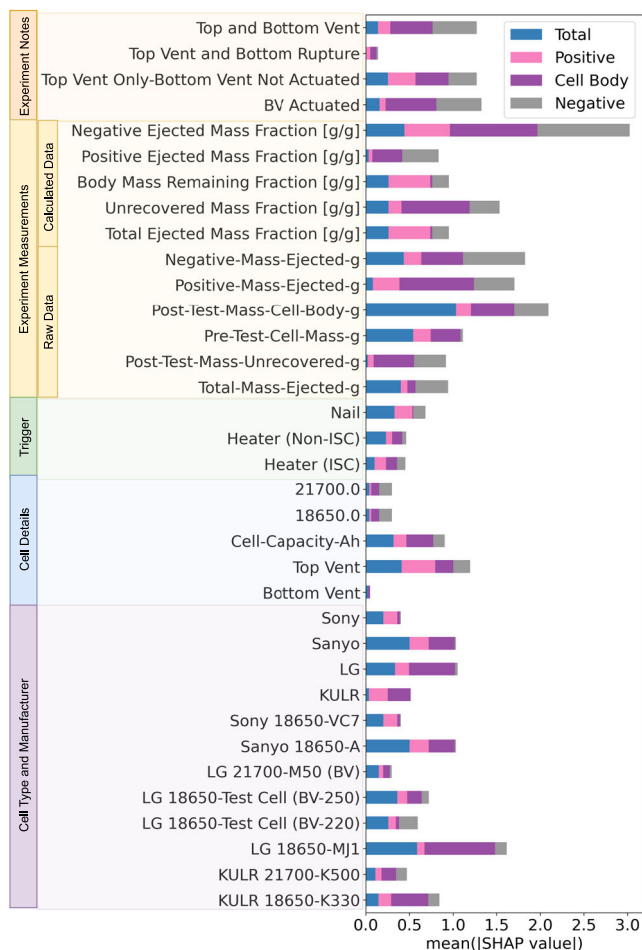
Fig. 9 | Mean absolute Shapley additive explanations (SHAP) value of each feature for each of the four model targets. Total heat output (blue), positive heat output (pink), cell body heat output (purple), and negative heat output (grey). A high SHAP value corresponds to a large impact on the model output. Features ordered into categories to ease interpretation.

Future work in battery safety will aim to extend the FTRC technique to other cell formats, like prismatic or pouch cells and, potentially, the approach shown here to estimate heat output from ejected mass could be extended to those new formats as well; given the safety risk and costs associated with destructive safety testing of larger cells, this would dramatically reduce costs of safety testing for large-format lithium-ion batteries. This approach should be combined with other types of battery safety tests, such as utilizing accelerating rate calorimetry to determine thermal runaway onset temperatures, to make quantitative safety maps of commercial lithium-ion batteries so that system engineers may better design zero-propagation battery packs for a variety of applications. The usefulness of the model demonstrated in this manuscript for estimating heat output of 18650 or 21700 format lithium-ion batteries should only improve as more FTRC data is collected and added to the Battery Failure Databank.

In summary, this work demonstrates that the distribution of fractional heat output during thermal runaway can be accurately predicted for many commercial 18650 and 21700 format lithium-ion batteries by leveraging a small number of fractional thermal runaway calorimetry (FTRC) measurements, supplemented by a large number of cheap and rapid measurements of ejected mass. Analysis of the performance of an SVM model, tested on 139 FTRC measurements from 8 different cell types, demonstrates that the average heat output of most cell types can be predicted with only a single FTRC

measurement with an error of less than $2kJ/Ah$ (about 10% error for the cells investigated here), and that only 5 FTRC measurements are required to confidently estimate the distribution of heat outputs including the occurrence of anomalously high heat outputs. This is achieved by using ML to estimate heat output from 10-30 ejected mass measurements. Adding further ejected mass measurements than available here would simply increase confidence that outlier values of the heat output distribution for any cell are adequately predicted. Additionally, the ML approach is able to distinguish cells with highly variable heat output with only 5 FTRC measurements, flagging out cells with highly variable failure from those with more predictable failures and prompting investment in further calorimetry measurements for 'unruly' cell types. Further testing on a secondary test set of cells with fewer than 10 measurements or with structural modifications demonstrates good performance of the trained SVM model when applied to new data, as well as demonstrating how the machine-learned model can be used to identify cells where ejected mass and heat output are not correlated, and thus additional FTRC measurements would be required to improve estimates of the heat output variability.

Implementing data driven methods could vastly reduce the amount of resources required for battery safety testing by predicting an initial distribution of heat output using few experimental calorimetry trials. However, the prediction approached used here is not recommended to replace experiments and instead only used as a preliminary estimating tool to evaluate the suitability of cells for specific applications. The current training data only covers pristine cells at full state of charge with a limited range of electrode and electrolyte chemistries; while the Battery Failure Databank continues to expand to include new cell chemistries, variable states of charge, and variable cycle histories, this approach can help improve confidence in the safety and reliability of lithium-ion batteries, by providing a method to accelerate the acquisition of critical data for designing safe lithium-ion battery systems.

Methods
Data collection
Experiments measuring the thermal runaway outcomes using a Fractional Thermal Runaway Calorimeter (FTRC)[5,8] were taken over many years and collected into the Battery Failure Databank[25]. Hundreds of thermal runaway events from various cell designs were measured by recording the total heat output, heat ejected from the positive end of the cell, heat ejected from the negative end, heat from the cell casing, as well as mass ejection from the cells. Most FTRC measurements were conducted at synchrotron facilities that facilitated simultaneous high-speed X-ray radiography to visualize internal occurrences during thermal runaway, but these data are not used in this work. The version of the Battery Failure Databank used here was V2 that was available open-access on November 2023. Additions to the Databank are expected to occur periodically over the following years.

Data pre-processing
A subset of the databank was used to maintain a balanced dataset for training models. Custom cells in the Battery Failure Databank, such as those with Soteria materials or test-cells with embedded Internal Short Circuiting Devices (ISCDs)[1], were excluded from the training set. Only cells with more than 10 samples were used. Only trials with commercial cells at 100% state of charge were included.

The total and fractional heat outputs were normalized by dividing by the discharge capacity of the cell (kJ/Ah). Prior exploration of trends in the Battery Failure Databank[1,2,4,5] revealed that normalizing heat output by capacity resulted in clear relationships between the heat output and other measured quantities such as the ejected mass fractions.

Modeling approach

Regression models predicted normalized total, positive, cell body, and negative heat outputs using a chain regression approach (predicted values from prior regressions are used as features for further regressions), on the basis that the total heat output is the sum of the fractional heat outputs. The chain regression was done in the order (cell body, negative, positive, total), as the cell body heat output was expected to be the easiest to predict, avoiding the accumulation of errors. It was found that model performance was not very sensitive to chain order. A risk of the chain regression approach is that errors will accumulate when making test predictions, but separate models treating each heat output as independent resulted in marginally worse performance across all cells. Regression targets were all z-score normalized prior on the training set.

Features for making predictions included cell metadata (cell type, manufacturer, format, charge capacity, TR trigger mechanism), FTRC experiment notes (cell failure mechanism, bottom vent actuation), and mass data after TR, in both absolute and normalized values (total ejected mass, positive ejected mass, negative ejected mass, cell body remaining mass, unaccounted mass, i.e., initial mass minus all measured masses after TR). The numerical features were all z-score normalized on the training set. Categorical features such as cell manufacturer were one-hot-encoded, i.e. converted to a form that can be interpreted by ML models by treating each category option as an independent variable with 1 or 0 binary values.

'Zero-shot' models performed predictions on test data from one held-out cell type at a time, making heat output predictions with no measured heat outputs from that cell. 'i-shot' models made predictions for held-out cell types after copying i samples from the test data to the training data. But, this introduces model bias due to the selected samples, for instance, a cell type with 30 FTRC measurements will have 30 separate '1-shot' models, each with their own predictions and error statistics. However, it is not feasible to sample every possible combination of samples as there are $\binom{n}{i}$ possible combinations of i cells from n total samples. So, i-shot models were either trained on 300 randomly sampled sets of sample combinations for $\binom{n}{i}$ greater than 300, or all combinations for $\binom{n}{i}$ less than or equal to 300, and error metrics like root-mean-square-error (RMSE) or Kullback-Leibler (KL) divergence are reported as distributions for each i. Figure 3 shows our process and evaluation.

Two model architectures were used, a baseline model and a Support Vector Machine (SVM) model with a linear kernel, squared L_2 norm weight of 1, and an epsilon of 0.1. The baseline model used just the mass features, performing a global linear regression for zero-shot predictions and a local linear regression on just the i samples copied from the held-out cell type for i-shot predictions. A SVM was chosen as the ML architecture, as it works well for small data sets and optimizes for the a hyperplane where the margin between the points closest to the decision boundary are maximized[29], which was hoped to result in accurate predictions of the heat output distribution. An XGBoost architecture was also tried, and resulted in slightly worse performance across all cell-types.

The model predictions were further evaluated by calculating the Mahalanobis depth for the predicted and actual values. The predicted and actual heat outputs were plotted against the ejected masses. The Mahalanobis depth is a statistical depth function that accounts for the distance one point is from all other points in a distribution considering the covariance of the data. The covariance matrix $\widehat{\sum_x^{-1}}$ describes how

close a point is from every other point in the data set.

$$D = \frac{1}{1 + (z - \bar{x})\widehat{\sum_x^{-1}}(z - \bar{x})} \tag{1}$$

A point with higher depth is more similar to other points in the distribution. The various depths may then be plotted using ellipses. The ellipse borders are cutoffs, where an ellipse contains some density of points.

The ML models used in this work were developed in Python using the sklearn library[30].

Data availability

The data used in this study is available open-access in the Battery Failure Databank: https://www.nrel.gov/transportation/battery-failure.html. Given that there may be future updates to the Battery Failure Databank, the version used in this work is also available from Zenodo[31]. Source data are provided with this paper.

Code availability

The code used in this work is available open-access at https://github.com/NREL/battery-heat-output and Zenodo[31].

References

1. Finegan, D. P. et al. Characterising thermal runaway within lithium-ion cells by inducing and monitoring internal short circuits. *Energy Environ. Sci.* **10**, 1377–1388 (2017).
2. Finegan, D. P. et al. Tracking internal temperature and structural dynamics during nail penetration of lithium-ion cells. *J. Electrochem. Soc.* **164**, A3285 (2017).
3. Sun, P., Bisschop, R., Niu, H. & Huang, X. A review of battery fires in electric vehicles. *Fire Technol.* **56**, 1361–1410 (2020).
4. Finegan, D. P. et al. Identifying the cause of rupture of li-ion batteries during thermal runaway. *Adv. Sci.* **5**, 1700369 (2018).
5. Walker, W. Q. et al. Decoupling of heat generated from ejected and non-ejected contents of 18650-format lithium-ion cells using statistical methods. *J. Power Sources* **415**, 207–218 (2019).
6. Walker, W. Q. et al. The effect of cell geometry and trigger method on the risks associated with thermal runaway of lithium-ion batteries. *J. Power Sources* **524**, 230645 (2022).
7. Sharp, M. et al. Thermal runaway of li-ion cells: How internal dynamics, mass ejection, and heat vary with cell geometry and abuse type. *J. Electrochem. Soc.* **169**, 020526 (2022).
8. Finegan, D. P. et al. Modelling and experiments to identify high-risk failure scenarios for testing the safety of lithium-ion cells. *J. Power Sources* **417**, 29–41 (2019).
9. Coman, P. T., Darcy, E. C., Strangways, B. & White, R. E. A reduced-order lumped model for li-ion battery packs during operation. *J. Electrochem. Soc.* **168**, 100525 (2021).
10. Coman, P. T., Darcy, E. C. & White, R. E. Simplified thermal runaway model for assisting the design of a novel safe li-ion battery pack. *J. Electrochem. Soc.* **169**, 040516 (2022).
11. Gao, S. et al. Experimental study on module-to-module thermal runaway-propagation in a battery pack. *J. Electrochem. Soc.* **166**, A2065 (2019).
12. Deng, J., Bae, C., Marcicki, J., Masias, A. & Miller, T. Safety modelling and testing of lithium-ion batteries in electrified vehicles. *Nat. Energy* **3**, 261–266 (2018).
13. Börger, A., Mertens, J. & Wenzl, H. Thermal runaway and thermal runaway propagation in batteries: What do we talk about? *J. Energy Storage* **24**, 100649 (2019).
14. Kim, J., Mallarapu, A., Finegan, D. P. & Santhanagopalan, S. Modeling cell venting and gas-phase reactions in 18650 lithium ion batteries during thermal runaway. *J. Power Sources* **489**, 229496 (2021).

15. Naha, A. et al. Internal short circuit detection in li-ion batteries using supervised machine learning. *Sci. Rep.* **10**, 1–10 (2020).
16. Chemali, E., Kollmeyer, P. J., Preindl, M. & Emadi, A. State-of-charge estimation of li-ion batteries using deep neural networks: A machine learning approach. *J. Power Sources* **400**, 242–255 (2018).
17. Sidhu, M. S., Ronanki, D. & Williamson, S. State of charge estimation of lithium-ion batteries using hybrid machine learning technique. *IECON 2019-45th Annu. Conf. IEEE Ind. Electron. Soc.* **1**, 2732–2737 (2019).
18. Fan, J., Fan, J., Liu, F., Qu, J. & Li, R. A novel machine learning method based approach for li-ion battery prognostic and health management. *IEEE Access* **7**, 160043–160061 (2019).
19. Gasper, P., Schiek, A., Smith, K., Shimonishi, Y. & Yoshida, S. Predicting battery capacity from impedance at varying temperature and state of charge using machine learning. *Cell Rep. Phys. Sci.* **3**, 101184 (2022).
20. Aykol, M. et al. Perspective—combining physics and machine learning to predict battery lifetime. *J. Electrochem. Soc.* **168**, 030525 (2021).
21. Kunz, M. R. et al. Early battery performance prediction for mixed use charging profiles using hierarchal machine learning. *Batteries Supercaps* **4**, 1186–1196 (2021).
22. Severson, K. A. et al. Data-driven prediction of battery cycle life before capacity degradation. *Nat. Energy* **4**, 383–391 (2019).
23. Yamanaka, T., Takagishi, Y. & Yamaue, T. A framework for optimal safety li-ion batteries design using physics-based models and machine learning approaches. *J. Electrochem. Soc.* **167**, 100516 (2020).
24. Takagishi, Y., Yamanaka, T. & Yamaue, T. Machine learning approaches for designing mesoscale structure of li-ion battery electrodes. *Batteries* **5**, 54 (2019).
25. Finegan, D. P. et al. The battery failure databank: Insights from an open-access database of thermal runaway behaviors of li-ion cells and a resource for benchmarking risks. *J. Power Sources* **597**, 234106 (2024).
26. NREL & NASA. Battery Failure Databank. https://www.nrel.gov/transportation/battery-failure.html (Accessed 10 December 2021).
27. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (2016).
28. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I. et al. (eds.) *Advances in Neural Information Processing Systems 30*, 4765–4774 (Curran Associates, Inc., 2017). http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
29. Keerthi, S. S., Shevade, S. K., Bhattacharyya, C. & Murthy, K. R. K. Improvements to platt's smo algorithm for svm classifier design. *Neural Comput.* **13**, 637–649 (2001).
30. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
31. Masalkovaite, K. & Gasper, P. Nrel/battery-heat-output: Publication (2024). https://doi.org/10.5281/zenodo.13125063

Acknowledgements

Author contributions

K.M. performed the data analysis, developed the method and code, and wrote the manuscript. P.G. contributed to data analysis, preparing the code for open-access sharing on a GitHub repository, provided guidance on the technique, and contributed to writing and review of the manuscript. D.P.F. facilitated the open-access Battery Failure Databank, provided insight and direction to the project, and contributed to writing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-52653-3.

Correspondence and requests for materials should be addressed to Paul Gasper or Donal P. Finegan.

Peer review information *Nature Communications* thanks Zhongbao Wei, and the other, anonymous, reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

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