
Quantifying cross-modal transfer in astronomical transient classification

Francisco Förster*

Data & Artificial Intelligence Initiative,
Universidad de Chile.

Daniel Moreno-Cartagena

Center for Data & Artificial Intelligence,
Universidad de Concepción.

Mathias González-Espinoza

Data & Artificial Intelligence Initiative,
Universidad de Chile.

Guillermo Cabrera-Vives

Center for Data & Artificial Intelligence,
Universidad de Concepción.

Pablo Estévez

Electrical Engineering Department,
Universidad de Chile.

Javier Correa-Orellana

Armagh Observatory and Planetarium,
Armagh, Northern Ireland.

Steve Jurado

Data & Artificial Intelligence Initiative,
Universidad de Chile.

Abstract

We investigate whether the classification of observed astronomical transients exhibits cross-modal transfer: the improved performance of models trained with more modalities than are available at inference. We consider models trained with combinations of light curves, images and spectra using different modality drop and fine-tuning strategies. For a sample of observed transients, we confirm that multimodal models show better macro F1 classification performance, with similar or marginally worse calibration errors. We find no significant cross-modal transfer for light-curve-only inference, but we detect a small, significant transfer for light-curve and image inference (+0.055 in macro F1, $p < 0.001$) when training with an imbalanced modality drop. Classic models trained with manually engineered features achieve better classification metrics than light curve-only inference models, but their raw scores are much less well calibrated.

1 Introduction

We perceive and understand the world in a multi-modal (MM) fashion, relying on complementary inputs such as vision, touch and hearing to navigate our environment. Motivated by this observation, MM machine learning (ML) models are increasingly used in a diversity of applications, ranging from chat bots to scientific applications. Some of the key ideas behind these models have been the representation, alignment, fusion and transference of modalities [Liang et al., 2024]. For example, via unsupervised training and contrastive loss alignment ([van den Oord et al., 2018, Radford et al., 2021]) and the trainable combination of modalities for supervised classification [Arevalo et al., 2017].

The implementation of MM models can be challenging. The alignment of modality representations often shows significant gaps even in well established MM models [Liang et al., 2022]. Uni-modal models can show similar performance to MM models in spite of having less input information due to

*fforster@uchile.cl, <https://fforster.github.io/>

several factors: 1) overfitting, as these models span a larger parameter space and the rate of overfitting can differ strongly between different modalities [Wang et al., 2020]; 2) gradient suppression, when gradients are dominated by one or a few modalities, effectively suppressing the gradients associated with the weaker modalities [Peng et al., 2022]; and 3) unimodal bias, where one modality significantly dominates the training process the later the modal fusion is performed [Zhang and Latham, 2024].

Several works have used MM ML in astronomy, from the study of galaxies combining images and spectra [Parker et al., 2024, Doorenbos et al., 2024] to time-domain astronomy combining light curves (LCs), images, spectra, and metadata from galactic and extragalactic variable phenomena. [Cabrera-Vives et al., 2024] combine synthetic LCs and metadata from variable sources using transformers [Vaswani et al., 2017] and late fusion. [Zhang et al., 2024] pre-train a model that combines synthetic supernova LCs and spectra using transformers and contrastive loss, later fine-tuned with real data from the Zwicky Transient Facility (ZTF, [Bellm et al., 2019]). [Sheng et al., 2024] combine real ZTF discovery images, early-time LCs and host galaxy information via convolutional neural networks and intermediate fusion. [Huijse et al., 2025] combine LCs and spectra via concatenation of variational autoencoder latent representations. [Rizhko and Bloom, 2025] combine real LCs, images and metadata from variable stars using three encoders and a contrastive loss. [Junell et al., 2026] combine real transient LCs, images, metadata and spectra using transformers, 2D and multi-scale 1D convolutional networks, and a mixture-of-experts fusion.

Our work focuses on time-domain astronomy and the case of real extragalactic transient objects, stellar explosions occurring in external galaxies, observed by ZTF. They are detected via image difference techniques and can be characterized by their multi-band photometric flux time series, or LCs; by their host galaxies, whose properties are correlated with their stellar progenitors; and by spectroscopic observations that reveal their underlying physical structure and composition. Consequently, we consider the use of LCs, contextual images and spectra for their classification.

While large astronomical surveys such as ZTF and the Legacy Survey of Space and Time (LSST, [Ivezić et al., 2019]) provide abundant sets of images and LCs, spectroscopic data is relatively rare and seldom available at inference time, even with new large spectroscopic surveys [DESI Collaboration et al., 2016, de Jong et al., 2019]. Thus, we are interested in the problem of cross-modal transfer, where a set of modalities used at training time can improve the internal representation of the model even when they are absent at inference time.

2 Data

We use a combination of LCs taken from ZTF via the ALerCE broker [Förster et al., 2021], host galaxy r-band multi-resolution images from PanSTARRS DR1 [Magnier et al., 2020]; and spectra from the Weizmann Interactive Supernova Data Repository (WiSeREP, [Yaron and Gal-Yam, 2012]). We use r-band multi-resolution images (as in [Förster et al., 2021]) instead of multi-band multi-resolution images because we rely on a local server that can match the ZTF and LSST alert rates using the r-band only, with the objective of putting these models in production.

We compile a sample of confirmed and candidate transient events to train in a semi-supervised/distilled fashion. For the confirmed transients, we use the classifications of the Transient Name Server (TNS, [IAU Supernova Working Group, 2016]) mapped into a simple taxonomy of supernovae (SNe) and related events: Type Ia SNe (SNIa), Stripped-Envelope SNe (SESN), Type II SNe (SNII), Superluminous SNe (SLSN) and Tidal Disruption Events (TDE). For the transient candidates, we select them cross-matching TNS with a catalog of probabilistic transient classifications from ALerCE, filtered according to their galactic latitude ($|b| > 10$) and data quality cuts. Our final catalog contains 99,317 astronomical events, of which 7,375 have a spectroscopically confirmed class. The class distribution is highly imbalanced; with 5,277 SNIa, 1,213 SNII, 450 SESN, 216 SNIIn, 158 SLSN, and 61 TDE. Modality-specific quality criteria are first applied to determine whether each modality contains usable information for a given object. The retained observations are subsequently transformed into fixed-size representations suitable for their corresponding encoders. Each modality preprocessing is described in Appendix C.

3 Methodology

3.1 Training set partition

We reserve 20% of the labeled sample ($N = 1,475$) as a held-out classification test set. Since our objective is to study cross-modal transfer; in particular, the effect of using spectroscopy during training while removing it at inference; the test set is restricted to objects with valid observations in all three modalities. The split is class-stratified, preserving the class distribution of the labeled sample. Consequently, every inference configuration is evaluated on the same objects. The remaining 5,900 labeled objects form a five-fold stratified cross-validation pool, with 4,720 training and 1,180 validation objects in each fold. Objects lacking one or more modalities are retained in this pool rather than discarded. The detailed class distribution among folds is shown in Appendix D.

For self-supervised pretraining, a separate partition is constructed for each modality using all objects that satisfy its quality criteria. We reserve 10% of each modality-specific dataset for reconstruction testing and 10% of the remaining objects for validation. To prevent leakage into downstream classification, the 1,475 classification-test objects are forced into the corresponding pretraining test partitions and never appear during pretraining or validation. Thus, the resulting train/validation/test partitions for self-supervised pretraining contain 59,335/6,593/7,325 objects for LCs, 13,823/1,536/1,707 for spectra, and 79,460/8,829/9,810 for images.

3.2 Architecture

Our framework starts with three modality-specific masked autoencoders (MAEs, [He et al., 2021]): for LCs, spectra (SPEC), and multi-resolution images (IM) as described in Appendix E.

For multimodal alignment, the three encoder outputs are pooled and mapped through modality-specific projection heads into a shared embedding space. The projected representations are aligned using a symmetric InfoNCE objective [van den Oord et al., 2018] over the available modality pairs (LC, IM), (LC, SPEC), and (IM, SPEC), with learnable inverse temperature. Only objects containing both modalities contribute to a given pairwise similarity matrix, and the objective is averaged over the modality pairs represented in the batch. Positive pairs correspond to different modalities of the same transient, allowing both labeled and unlabeled objects to contribute to the loss function.

For multimodal classification, the projected representations are fused using gated multimodal unit (GMU, [Arevalo et al., 2017]). Unavailable modality embeddings and their availability flags are set to zero, allowing the GMU to adapt its weighting to the information present for each object. The fused representation is passed to a Multi-layer Perceptron classifier, producing six class logits.

3.3 Training

Training proceeds sequentially through unimodal self-supervised pretraining, multimodal contrastive alignment, and supervised classification. Hyperparameters are optimized independently at each stage using validation data only; the held-out test set is never used for model or hyperparameter selection.

We train eight different classifiers: a LC-only classifier that uses the LC encoder and where the supervised classification only sees LCs; a LC+IM classifier that uses the LC and IM encoders and where alignment, fusion and supervised classification do not see spectra; and six different LC+IM+SPEC classifiers that use the three encoders and where the alignment, fusion and supervised classification can see all data modalities, but differ on how modalities are dropped during supervised classification and on whether we fine-tune the model for inference with specific modalities. One LC+IM+SPEC model is trained dropping modalities in a balanced fashion during supervised classification (25% of the time no modalities are dropped, 25% of the time LCs are dropped, 25% of the time IMs are dropped, and 25% of the time SPECs are dropped). Another LC+IM+SPEC model is trained dropping modalities in an imbalanced fashion (68% of the time they see only LCs, 29% of the time they see LCs and IMs, 3% of the time they see LCs and SPEC). Then, for the cross-modal transfer case, the two previous models are fine-tuned using only LCs and using both LCs and IMs during training. Note that we always keep unfrozen weights for the alignment and fusion stages when doing supervised classification training. More information about the training process is discussed in Appendix F.

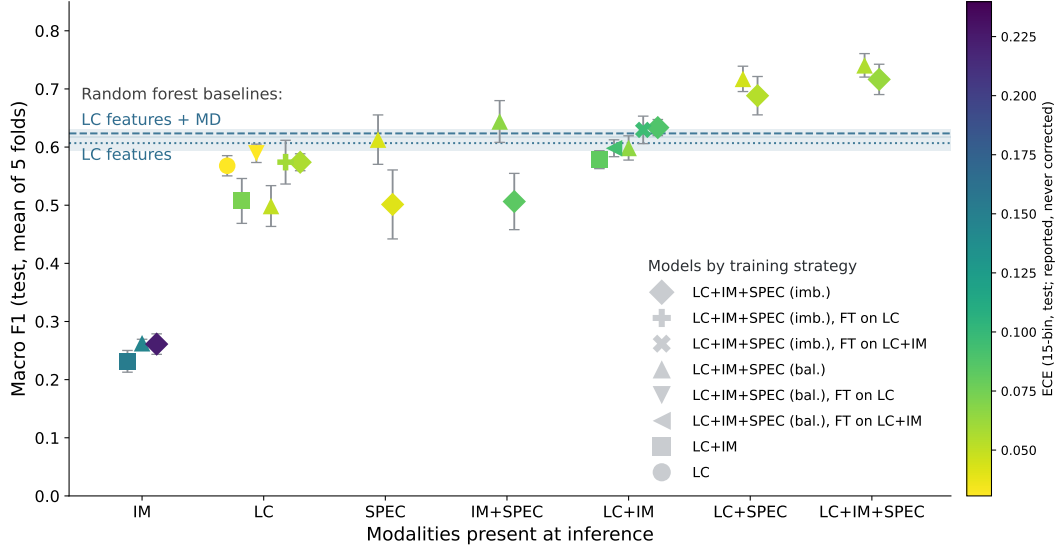


Figure 1: Mean macro F1 scores and one standard deviation error bars, colored by ECE, for models using different training modalities and training strategies as a function of inference modalities. The baseline models are shown for reference and are also colored by ECE. See detailed values in Table 4.

3.4 Feature-based baseline

To provide a classical reference for the learned representations, we train two random forest (RF) classifiers on extensively tested, manually engineered features computed with the ALerCE ZTF feature extractor. The complete extractor produces 209 quantities spanning statistical variability descriptors, colors, harmonic and period-based statistics, multi-harmonic power-spectrum features, and parameters from several transient-specific models. The complete list is provided in Appendix G.

We consider two feature configurations. The *LC-only* baseline contains 158 features derived exclusively from the LCs and provides a direct comparison with models receiving the same modality. The *LC+metadata* baseline contains 178 features and additionally incorporates contextual information from AllWISE [Cutri et al., 2021], Pan-STARRS, and ZTF reference-image measurements. This second configuration quantifies the contribution of catalog-level context beyond the LC itself.

Both RF baselines use class-balanced weights and undergo independent hyperparameter optimization using an Optuna Tree-structured Parzen Estimator (TPE) sampler, as reported in Appendix F.

3.5 Statistical comparison

All folds share one test set and $\sim 75\%$ of their training data, so the fold-to-fold standard deviation measures mainly retraining variability. To test cross-modal transfer we compare each model with the model trained on the inference modalities alone using a paired bootstrap over the 1,475 test objects (10^4 resamples): every resample is applied to both models and to all five fold models, and we record the difference of their fold-averaged macro F1.

4 Results

We evaluate the previous models on exactly the same held-out test set using different combinations of modalities during inference and report the macro F1 and expected calibration error (ECE) statistics in Figure 1, where the baseline RF classifiers macro F1 and ECE are also shown. A detailed table with macro F1 and ECE statistics, confusion matrices and reliability curves are shown in Appendix H. All scores are reported raw: we apply no post-hoc calibration to any model.

It can be seen that the best macro F1 statistics are obtained for the models that use all three modalities during inference, with marginally better statistics using a balanced modality drop strategy during

training. Marginally better ECE metrics are found for selected LC-only inference models. It can also be seen that the most informative modalities appear to follow the domain intuition: SPEC > LC > IM.

The classic RF baseline models have equal or better macro F1 than all LC-only inference models considered, although with much higher raw ECE (0.165–0.166 versus 0.031–0.073). Their higher macro F1 may be due to a combination of several effects: the small sample size, where neural networks are sometimes outperformed by classical methods; the highly imbalanced nature of the dataset, where in our experience class weighted RFs can provide very good results; and the strong inductive bias introduced by the manually engineered features. The contribution of these features is confirmed by the suboptimal performance of the LC-only models in [Cabrera-Vives et al., 2024]. Their reliability curves show underconfidence (Figure 4), as expected for random forests [Niculescu-Mizil and Caruana, 2005], and which class-balanced weighting can accentuate. Among the neural models, the multimodal ones are overconfident at their native input relative to the LC-only and LC+IM models, an effect reduced by fine-tuning on LC only (Figure 4).

Regarding cross-modal transfer, the main question in this work, we find none for LC-only inference: no model trained with additional modalities significantly outperforms the LC-only model (Table 5). For LC+IM inference, the LC+IM+SPEC model trained with an imbalanced modality drop outperforms the LC+IM model by $\Delta F_1 = +0.055$ (95% CI [+0.026, +0.087], $p < 0.001$), and by +0.051 [+0.024, +0.081] after fine-tuning on LC+IM, with a positive gain in all five folds. Balanced modality drop yields no significant transfer.

The two drop strategies expose the classifier to different input patterns. The balanced strategy removes at most one modality at a time, so LC-only inputs are rare in training (about 2%, from objects without a spectrum), and spectra are present in about 70% of examples. The imbalanced strategy shows LC-only inputs 68% of the time, LC+IM 29% of the time, and includes spectra only in the remaining 3% (as LC+SPEC). Accordingly, the balanced model is best whenever spectra are available at inference (Table 4), but falls below the LC-only model when evaluated on LC alone ($\Delta F_1 = -0.069$, Table 5), a gap that fine-tuning on LC closes. The imbalanced model, trained mostly without spectra, learns stronger LC and LC+IM pathways, which is where its transfer appears.

5 Conclusions

We confirm the importance of building LC+IM+SPEC classification models for transient classification, with trimodal inference models outperforming all LC and LC+IM inference models in macro F1. LC-only inference models could not outperform classic RF models in macro F1, but had better raw ECE. We detect no significant cross-modal transfer for the LC-only inference case, but we detect cross-modal transfer in the LC+IM inference case ($\Delta F_1 = +0.055$, $p < 0.001$). These estimates come from a small and imbalanced labeled sample: the rarest classes have 12 (TDE) and 32 (SLSN) test objects, which dominate the macro F1 uncertainty, and may explain why cross-modal transfer was small. This question remains open for LSST, with significantly larger samples, where access to spectra will be limited, and where labeled sets will be strongly biased to brighter and more nearby objects.

References

- J. Arevalo, T. Solorio, M. M. y Gómez, and F. A. González. Gated multimodal units for information fusion. *ArXiv*, abs/1702.01992, 2017. URL <https://api.semanticscholar.org/CorpusID:9401721>.
- E. C. Bellm, S. R. Kulkarni, M. J. Graham, R. Dekany, R. M. Smith, R. Riddle, F. J. Masci, G. Helou, T. A. Prince, S. M. Adams, C. Barbarino, T. Barlow, J. Bauer, R. Beck, J. Belicki, R. Biswas, N. Blagorodnova, D. Bodewits, B. Bolin, V. Brinnel, T. Brooke, B. Bue, M. Bulla, R. Burruss, S. B. Cenko, C.-K. Chang, A. Connolly, M. Coughlin, J. Cromer, V. Cunningham, K. De, A. Delacroix, V. Desai, D. A. Duev, G. Eadie, T. L. Farnham, M. Feeney, U. Feindt, D. Flynn, A. Franckowiak, S. Frederick, C. Fremling, A. Gal-Yam, S. Gezari, M. Giomi, D. A. Goldstein, V. Z. Golkhou, A. Goobar, S. Groom, E. Hacopians, D. Hale, J. Henning, A. Y. Q. Ho, D. Hover, J. Howell, T. Hung, D. Huppenkothen, D. Imel, W.-H. Ip, Ž. Ivezić, E. Jackson, L. Jones, M. Juric, M. M. Kasliwal, S. Kaspi, S. Kaye, M. S. P. Kelley, M. Kowalski, E. Kramer, T. Kupfer, W. Landry, R. R. Laher, C.-D. Lee, H. W. Lin, Z.-Y. Lin, R. Lunnan, M. Giomi, A. Mahabal, P. Mao, A. A.

- Miller, S. Monkewitz, P. Murphy, C.-C. Ngeow, J. Nordin, P. Nugent, E. Ofek, M. T. Patterson, B. Penprase, M. Porter, L. Rauch, U. Rebbapragada, D. Reiley, M. Rigault, H. Rodriguez, J. van Roestel, B. Rusholme, J. van Santen, S. Schulze, D. L. Shupe, L. P. Singer, M. T. Soumagnac, R. Stein, J. Surace, J. Sollerman, P. Szkody, F. Taddia, S. Terek, A. Van Sistine, S. van Velzen, W. T. Vestrand, R. Walters, C. Ward, Q.-Z. Ye, P.-C. Yu, L. Yan, and J. Zolkower. The Zwicky Transient Facility: System Overview, Performance, and First Results. *Publications of the Astronomical Society of the Pacific*, 131(995):018002, Jan. 2019. doi: 10.1088/1538-3873/aaecbe.
- G. Cabrera-Vives, D. Moreno-Cartagena, N. Astorga, I. Reyes-Jainaga, F. Förster, P. Huijse, J. Arredondo, A. M. Muñoz Arancibia, A. Bayo, M. Catelan, P. A. Estévez, P. Sánchez-Sáez, A. Álvarez, P. Castellanos, P. Gallardo, A. Moya, and D. Rodriguez-Mancini. ATAT: Astronomical Transformer for time series and Tabular data. *A&A*, 689:A289, Sept. 2024. doi: 10.1051/0004-6361/202449475.
- R. M. Cutri, E. L. Wright, T. Conrow, J. W. Fowler, P. R. M. Eisenhardt, C. Grillmair, J. D. Kirkpatrick, F. Masci, H. L. McCallon, S. L. Wheelock, S. Fajardo-Acosta, L. Yan, D. Benford, M. Harbut, T. Jarrett, S. Lake, D. Leisawitz, M. E. Ressler, S. A. Stanford, C.-W. Tsai, F. Liu, G. Helou, A. Mainzer, D. Gettngs, A. Gonzalez, D. Hoffman, K. A. Marsh, D. Padgett, M. F. Skrutskie, R. Beck, M. Papin, and M. Wittman. VizieR Online Data Catalog: AllWISE Data Release (Cutri+2013). VizieR On-line Data Catalog: II/328. Originally published in: IPAC/Caltech (2013), Feb. 2021.
- R. S. de Jong, O. Agertz, A. A. Berbel, J. Aird, D. A. Alexander, A. Amarsi, F. Anders, R. Andrae, B. Ansarinejad, W. Ansorge, P. Antilogus, H. Anwand-Heerwart, A. Arentsen, A. Arnadottir, M. Asplund, M. Auger, N. Azais, D. Baade, G. Baker, S. Baker, E. Balbinot, I. K. Baldry, M. Banerji, S. Barden, P. Barklem, E. Barthélemy-Mazot, C. Battistini, S. Bauer, C. P. M. Bell, O. Bellido-Tirado, S. Bellstedt, V. Belokurov, T. Bensby, M. Bergemann, J. M. Bestenlehner, R. Bielby, M. Bilicki, C. Blake, J. Bland-Hawthorn, C. Boeche, W. Bolland, T. Boller, S. Bongard, A. Bongiorno, P. Bonifacio, D. Boudon, D. Brooks, M. J. I. Brown, R. Brown, M. Brüggen, J. Brynnel, J. Brzeski, T. Buchert, P. Buschkamp, E. Caffau, P. Caillier, J. Carrick, L. Casagrande, S. Case, A. Casey, I. Cesarini, G. Cescutti, D. Chapuis, C. Chiappini, M. Childress, N. Christlieb, R. Church, M.-R. L. Cioni, M. Cluver, M. Colless, T. Collett, J. Comparat, A. Cooper, W. Couch, F. Courbin, S. Croom, D. Croton, E. Daguisé, G. Dalton, L. J. M. Davies, T. Davis, P. de Laverny, A. Deason, F. Dionies, K. Disseau, P. Doel, D. Döscher, S. P. Driver, T. Dwelly, D. Eckert, A. Edge, B. Edvardsson, D. E. Youssoufi, A. Elhaddad, H. Enke, G. Erfanianfar, T. Farrell, T. Fechner, C. Feiz, S. Feltzing, I. Ferreras, D. Feuerstein, D. Feuillet, A. Finoguenov, D. Ford, S. Fotopoulou, M. Fouesneau, C. Frenk, S. Frey, W. Gaessler, S. Geier, N. Gentile Fusillo, O. Gerhard, T. Giannantonio, D. Giannone, B. Gibson, P. Gillingham, C. González-Fernández, E. Gonzalez-Solares, S. Gottloeber, A. Gould, E. K. Grebel, A. Gueguen, G. Guiglion, M. Haehnelt, T. Hahn, C. J. Hansen, H. Hartman, K. Hauptner, K. Hawkins, D. Haynes, R. Haynes, U. Heiter, A. Helmi, C. H. Aguayo, P. Hewett, S. Hinton, D. Hobbs, S. Hoenig, D. Hofman, I. Hook, J. Hopgood, A. Hopkins, A. Hourihane, L. Howes, C. Howlett, T. Huet, M. Irwin, O. Iwert, P. Jablonka, T. Jahn, K. Jahnke, A. Jarno, S. Jin, P. Jofre, D. Johl, D. Jones, H. Jönsson, C. Jordan, I. Karovicova, A. Khalatyan, A. Kelz, R. Kennicutt, D. King, F. Kitaura, J. Klar, U. Klauser, J.-P. Kneib, A. Koch, S. Koposov, G. Kordopatis, A. Korn, J. Kosmowski, R. Kotak, M. Kovalev, K. Kreckel, Y. Kripak, M. Krumpe, K. Kuijken, A. Kunder, I. Kushniruk, M. I. Lam, G. Lamer, F. Laurent, J. Lawrence, M. Lehmitz, B. Lemasle, J. Lewis, B. Li, C. Lidman, K. Lind, J. Liske, J.-L. Lizon, J. Loveday, H.-G. Ludwig, R. M. McDermid, K. Maguire, V. Mainieri, S. Mali, and H. Mandel. 4MOST: Project overview and information for the First Call for Proposals. *The Messenger*, 175:3–11, Mar. 2019. doi: 10.18727/0722-6691/5117.
- DESI Collaboration, A. Aghamousa, J. Aguilar, S. Ahlen, S. Alam, L. E. Allen, C. Allende Prieto, J. Annis, S. Bailey, C. Balland, O. Ballester, C. Baltay, L. Beaufore, C. Bebek, T. C. Beers, E. F. Bell, J. L. Bernal, R. Besuner, F. Beutler, C. Blake, H. Bleuler, M. Blomqvist, R. Blum, A. S. Bolton, C. Briceno, D. Brooks, J. R. Brownstein, E. Buckley-Geer, A. Burden, E. Burtin, N. G. Busca, R. N. Cahn, Y.-C. Cai, L. Cardiel-Sas, R. G. Carlberg, P.-H. Carton, R. Casas, F. J. Castander, J. L. Cervantes-Cota, T. M. Claybaugh, M. Close, C. T. Coker, S. Cole, J. Comparat, A. P. Cooper, M.-C. Cousinou, M. Crocce, J.-G. Cuby, D. P. Cunningham, T. M. Davis, K. S. Dawson, A. de la Macorra, J. De Vicente, T. Delubac, M. Derwent, A. Dey, G. Dhungana, Z. Ding, P. Doel, Y. T. Duan, A. Ealet, J. Edelstein, S. Eftekharzadeh, D. J. Eisenstein, A. Elliott, S. Escoffier,

- M. Evatt, P. Fagrellius, X. Fan, K. Fanning, A. Farahi, J. Farihi, G. Favole, Y. Feng, E. Fernandez, J. R. Findlay, D. P. Finkbeiner, M. J. Fitzpatrick, B. Flaugher, S. Flender, A. Font-Ribera, J. E. Forero-Romero, P. Fosalba, C. S. Frenk, M. Fumagalli, B. T. Gaensicke, G. Gallo, J. Garcia-Bellido, E. Gaztanaga, N. Pietro Gentile Fusillo, T. Gerard, I. Gershkovich, T. Giannantonio, D. Gillet, G. Gonzalez-de-Rivera, V. Gonzalez-Perez, S. Gott, O. Graur, G. Gutierrez, J. Guy, S. Habib, H. Heetderks, I. Heetderks, K. Heitmann, W. A. Hellwing, D. A. Herrera, S. Ho, S. Holland, K. Honscheid, E. Huff, T. A. Hutchinson, D. Huterer, H. S. Hwang, J. M. Illa Laguna, Y. Ishikawa, D. Jacobs, N. Jeffrey, P. Jelinsky, E. Jennings, L. Jiang, J. Jimenez, J. Johnson, R. Joyce, E. Jullo, S. Juneau, S. Kama, A. Karcher, S. Karkar, R. Kehoe, N. Kennamer, S. Kent, M. Kilbinger, A. G. Kim, D. Kirkby, T. Kisner, E. Kitanidis, J.-P. Kneib, S. Koposov, E. Kovacs, K. Koyama, A. Kremin, R. Kron, L. Kronig, A. Kueter-Young, C. G. Lacey, R. Lafever, O. Lahav, A. Lambert, M. Lampton, M. Landriau, D. Lang, T. R. Lauer, J.-M. Le Goff, L. Le Guillou, A. Le Van Suu, J. H. Lee, S.-J. Lee, D. Leitner, M. Lesser, M. E. Levi, B. L'Huillier, B. Li, M. Liang, H. Lin, E. Linder, S. R. Loebman, Z. Lukić, J. Ma, N. MacCrann, C. Magneville, L. Makarem, M. Manera, C. J. Manser, R. Marshall, P. Martini, R. Massey, T. Matheson, J. McCauley, P. McDonald, I. D. McGreer, A. Meisner, N. Metcalfe, T. N. Miller, R. Miquel, J. Moustakas, A. Myers, M. Naik, J. A. Newman, R. C. Nichol, A. Nicola, L. Nicolati da Costa, J. Nie, G. Niz, P. Norberg, B. Nord, D. Norman, P. Nugent, T. O'Brien, M. Oh, and K. A. G. Olsen. The DESI Experiment Part I: Science, Targeting, and Survey Design. *arXiv e-prints*, art. arXiv:1611.00036, Oct. 2016. doi: 10.48550/arXiv.1611.00036.
- L. Doorenbos, E. Sestl, K. Heng, S. Cuvuoti, M. Brescia, O. Torbaniuk, G. Longo, R. Sznitman, and P. Márquez-Neila. Galaxy Spectroscopy without Spectra: Galaxy Properties from Photometric Images with Conditional Diffusion Models. *The Astrophysical Journal*, 977(1):131, Dec. 2024. doi: 10.3847/1538-4357/ad8bbe.
- F. Förster, G. Cabrera-Vives, E. Castillo-Navarrete, P. A. Estévez, P. Sánchez-Sáez, J. Arredondo, F. E. Bauer, R. Carrasco-Davis, M. Catelan, F. Elorrieta, S. Eyheramendy, P. Huijse, G. Pignata, E. Reyes, I. Reyes, D. Rodríguez-Mancini, D. Ruz-Mieres, C. Valenzuela, I. Álvarez Maldonado, N. Astorga, J. Borissova, A. Clocchiatti, D. De Cicco, C. Donoso-Oliva, L. Hernández-García, M. J. Graham, A. Jordán, R. Kurtev, A. Mahabal, J. C. Maureira, A. Muñoz-Arancibia, R. Molina-Ferreiro, A. Moya, W. Palma, M. Pérez-Carrasco, P. Protopapas, M. Romero, L. Sabatini-Gacitua, A. Sánchez, J. S. Martín, C. Sepúlveda-Cobo, E. Vera, and J. R. Vergara. The automatic learning for the rapid classification of events (alerce) alert broker. *The Astronomical Journal*, 161(5):242, apr 2021. doi: 10.3847/1538-3881/abe9bc. URL <https://doi.org/10.3847/1538-3881/abe9bc>.
- K. He, X. Chen, S. Xie, Y. Li, P. Doll'ar, and R. B. Girshick. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2021. URL <https://api.semanticscholar.org/CorpusID:243985980>.
- P. Huijse, J. De Ridder, L. Eyer, L. Rimoldini, B. Holl, N. Chornay, J. Roquette, K. Nienartowicz, G. Jevardat de Fombelle, D. J. Fritzewski, A. Kemp, V. Vanlaer, M. Vanrespaille, H. Wang, M. I. Carnerero, C. M. Raiteri, G. Marton, M. Madarász, G. Clementini, P. Gavras, and C. Aerts. Learning novel representations of variable sources from multi-modal gaia data via autoencoders. *Astronomy & Astrophysics*, 701:A150, 2025. doi: 10.1051/0004-6361/202554025. URL <https://aanda.org>.
- IAU Supernova Working Group. Transient name server. The IAU official mechanism for reporting new astronomical transients, 2016. URL <https://www.wis-tns.org/>.
- Ž. Ivezić, S. M. Kahn, J. A. Tyson, B. Abel, E. Acosta, R. Allsman, D. Alonso, Y. AlSayyad, S. F. Anderson, J. Andrew, J. R. P. Angel, G. Z. Angeli, R. Ansari, P. Antilogus, C. Araujo, R. Armstrong, K. T. Arndt, P. Astier, É. Aubourg, N. Auza, T. S. Axelrod, D. J. Bard, J. D. Barr, A. Barrau, J. G. Bartlett, A. E. Bauer, B. J. Bauman, S. Baumont, E. Bechtol, K. Bechtol, A. C. Becker, J. Becla, C. Beldica, S. Bellavia, F. B. Bianco, R. Biswas, G. Blanc, J. Blazek, R. D. Blandford, J. S. Bloom, J. Bogart, T. W. Bond, M. T. Booth, A. W. Borgland, K. Borne, J. F. Bosch, D. Boutigny, C. A. Brackett, A. Bradshaw, W. N. Brandt, M. E. Brown, J. S. Bullock, P. Burchat, D. L. Burke, G. Cagnoli, D. Calabrese, S. Callahan, A. L. Callen, J. L. Carlin, E. L. Carlson, S. Chandrasekharan, G. Charles-Emerson, S. Chesley, E. C. Cheu, H.-F. Chiang, J. Chiang, C. Chirino, D. Chow, D. R. Ciardi, C. F. Claver, J. Cohen-Tanugi, J. J. Cockrum, R. Coles, A. J.

- Connolly, K. H. Cook, A. Cooray, K. R. Covey, C. Cribbs, W. Cui, R. Cutri, P. N. Daly, S. F. Daniel, F. Daruich, G. Daubard, G. Daues, W. Dawson, F. Delgado, A. Dellapenna, R. de Peyster, M. de Val-Borro, S. W. Digel, P. Doherty, R. Dubois, G. P. Dubois-Felsmann, J. Durech, F. Economou, T. Eifler, M. Eracleous, B. L. Emmons, A. Fausti Neto, H. Ferguson, E. Figueroa, M. Fisher-Levine, W. Focke, M. D. Foss, J. Frank, M. D. Freemon, E. Gangler, E. Gawiser, J. C. Geary, P. Gee, M. Geha, C. J. B. Gessner, R. R. Gibson, D. K. Gilmore, T. Glanzman, W. Glick, T. Goldina, D. A. Goldstein, I. Goodenow, M. L. Graham, W. J. Gressler, P. Gris, L. P. Guy, A. Guyonnet, G. Haller, R. Harris, P. A. Hascall, J. Haupt, F. Hernandez, S. Herrmann, E. Hileman, J. Hoblitt, J. A. Hodgson, C. Hogan, J. D. Howard, D. Huang, M. E. Huffer, P. Ingraham, W. R. Innes, S. H. Jacoby, B. Jain, F. Jammes, M. J. Jee, T. Jenness, G. Jernigan, D. Jevremović, K. Johns, A. S. Johnson, M. W. G. Johnson, R. L. Jones, C. Juramy-Gilles, M. Jurić, J. S. Kalirai, N. J. Kallivayalil, B. Kalmbach, J. P. Kantor, P. Karst, M. M. Kasliwal, H. Kelly, R. Kessler, V. Kinnison, D. Kirkby, L. Knox, I. V. Kotov, V. L. Krabbendam, K. S. Krughoff, P. Kubánek, J. Kuczewski, S. Kulkarni, J. Ku, N. R. Kurita, C. S. Lage, R. Lambert, T. Lange, J. B. Langton, L. Le Guillou, D. Levine, M. Liang, K.-T. Lim, C. J. Lintott, K. E. Long, M. Lopez, P. J. Lotz, R. H. Lupton, N. B. Lust, L. A. MacArthur, A. Mahabal, R. Mandelbaum, T. W. Markiewicz, D. S. Marsh, P. J. Marshall, S. Marshall, M. May, R. McKercher, M. McQueen, J. Meyers, M. Migliore, M. Miller, and D. J. Mills. LSST: From Science Drivers to Reference Design and Anticipated Data Products. *The Astrophysical Journal*, 873(2):111, Mar. 2019. doi: 10.3847/1538-4357/ab042c.
- A. Junell, A. Sasli, F. F. Nunes, M. Xu, B. Border, N. Rehemtulla, M. Rizhko, Y.-J. Qin, T. Jegou Du Laz, A. Le Calloch, S. S. Chaudhary, S. Wu, J. Sollerman, N. Sravan, S. L. Groom, D. Hale, M. M. Kasliwal, J. Purdum, A. Wold, M. J. Graham, and M. W. Coughlin. Applying Multimodal Learning to Classify Transient Detections Early (AppleCiDer). I. Dataset, Methods, and Infrastructure. *Publications of the Astronomical Society of the Pacific*, 138(5):054508, May 2026. doi: 10.1088/1538-3873/ae55cb.
- P. P. Liang, A. Zadeh, and L.-P. Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Comput. Surv.*, 56(10), June 2024. ISSN 0360-0300. doi: 10.1145/3656580. URL <https://doi.org/10.1145/3656580>.
- W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Zou. Mind the gap: understanding the modality gap in multi-modal contrastive representation learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.
- E. A. Magnier, K. C. Chambers, H. A. Flewelling, J. C. Hoblitt, M. E. Huber, P. A. Price, W. E. Sweeney, C. Z. Waters, L. Denneau, P. W. Draper, K. W. Hodapp, R. Jedicke, N. Kaiser, R.-P. Kudritzki, N. Metcalfe, C. W. Stubbs, and R. J. Wainscoat. The Pan-STARRS Data-processing System. *The Astrophysical Journal Supplement Series*, 251(1):3, Nov. 2020. doi: 10.3847/1538-4365/abb829.
- A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning, ICML '05*, page 625–632, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102430. URL <https://doi.org/10.1145/1102351.1102430>.
- L. Parker, F. Lanusse, S. Golkar, L. Sarra, M. Cranmer, A. Bietti, M. Eickenberg, G. Krawezik, M. McCabe, R. Morel, R. Ohana, M. Pettee, B. Régalo-Saint Blancard, K. Cho, S. Ho, and Polymathic AI Collaboration. AstroCLIP: a cross-modal foundation model for galaxies. *Monthly Notices of the Royal Astronomical Society*, 531(4):4990–5011, July 2024. doi: 10.1093/mnras/stae1450.
- X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu. Balanced multimodal learning via on-the-fly gradient modulation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8228–8237, 2022. URL <https://api.semanticscholar.org/CorpusID:247779156>.
- A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/radford21a.html>.

- M. Rizhko and J. S. Bloom. AstroM³: A Self-supervised Multimodal Model for Astronomy. *The Astrophysical Journal*, 170(1):28, July 2025. doi: 10.3847/1538-3881/adcbad.
- X. Sheng, M. Nicholl, K. W. Smith, D. R. Young, R. D. Williams, H. F. Stevance, S. J. Smartt, S. Srivastav, and T. Moore. NEural Engine for Discovering Luminous Events (NEEDLE): identifying rare transient candidates in real time from host galaxy images. *Monthly Notices of the Royal Astronomical Society*, 531(2):2474–2492, June 2024. doi: 10.1093/mnras/stae1253.
- A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL <http://dblp.uni-trier.de/db/journals/corr/corr1807.html#abs-1807-03748>.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.
- W. Wang, D. Tran, and M. Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- O. Yaron and A. Gal-Yam. WISEREP—An Interactive Supernova Data Repository. *Publications of the Astronomical Society of the Pacific*, 124(917):668, July 2012. doi: 10.1086/666656.
- G. Zhang, T. Helfer, A. T. Gagliano, S. Mishra-Sharma, and V. Ashley Villar. Maven: a multimodal foundation model for supernova science. *Machine Learning: Science and Technology*, 5(4):045069, Dec. 2024. doi: 10.1088/2632-2153/ad990d.
- Y. Zhang and P. E. Latham. Understanding unimodal bias in multimodal deep linear networks. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 59231–59265. PMLR, 07–13 Jul 2024. URL <https://proceedings.mlr.press/v235/zhang24aa.html>.

A Acknowledgments

We thank ANID for infrastructure projects QUIMAL190012 and QUIMAL240008. F.F., P.A.E., G.C.V., S.J. thank Millennium Science Initiative Program ICN12_009, ANID Basal Program for Centers of Excellence FB210005, FONDECYT Regular 1200495, 1231877, 1241005, and 1262361. This research made use of services provided by the Centre de Données astronomiques de Strasbourg (CDS).

B Use of generative artificial intelligence (AI)

The core code used in the project was developed manually, but we implemented several modifications with the help of AI. We also used AI to streamline the experiment runs and visualize their results, monitoring relevant metrics via MLFlow and Wandb. The main scientific questions, experimental design, choice of evaluation metrics, and interpretation of the results were determined by the authors. We did not use AI tools to write the manuscript.

C Data quality and preprocessing

Light curves. LCs are cleaned independently in each band. Detections with a real–bogus score below $rb = 0.55$ and measurements with photometric uncertainties above the 95th percentile of σ_{psf} are removed. Observations sharing the same MJD to four decimal places are consolidated by taking their median. We additionally remove isolated observations at the temporal boundaries when the adjacent epoch lies in the extreme tail of the time distribution and its separation from the nearest neighbor exceeds the 97th percentile of consecutive temporal gaps. This procedure is applied

iteratively. A light curve is retained when at least eight valid observations remain in total across the g and r bands.

Observation times are expressed in days relative to a common reference time t_0 shared by both bands, preserving their relative temporal alignment. Each band is padded or truncated to a maximum of 65 observations and is accompanied by a per-band binary mask. A band containing no valid observations remains entirely masked.

Images. An image is retained when all five spatial scales are available. These scales are stacked into a $5 \times 30 \times 30$ tensor, providing the encoder with information over multiple spatial scales. Non-finite pixel values are replaced by zero.

Spectra. A spectrum is retained if it contains ten or more finite flux measurements within 3500–9500 Å. When multiple spectra satisfy this criterion, a single spectrum is selected by prioritizing wavelength coverage within this range, followed by signal-to-noise ratio and the number of sampled wavelength points.

The selected spectrum is de-redshifted with the catalog redshift and resampled onto a common rest-frame logarithmic wavelength grid of 750 bins spanning 3500–9500 Å, corresponding to a constant velocity spacing of 400 km s^{-1} . A moving average at the same velocity scale is used for local smoothing. The flux is shifted to non-negative values before estimating the continuum with a broader $30,000 \text{ km s}^{-1}$ smoothing window. Following the AstroDASH convention, the continuum-normalized spectrum is

$$f_{\text{norm}}(\lambda) = \frac{f(\lambda)}{f_{\text{cont}}(\lambda)} - 1, \quad (1)$$

which suppresses the slowly varying continuum while preserving spectral features. A cosine taper is applied over the outer 5% of the covered wavelength region to reduce edge discontinuities. Spectra of objects without a catalog redshift (103 labeled objects, 1.4%) are not corrected and remain in the observer frame. A binary mask identifies wavelength bins covered by the original observation. The median spectral coverage is 669 of the 750 bins.

D Class distributions

Table 1: Distribution of the labeled classes and unlabeled data modalities. Train and validation counts are for fold 0 of the five stratified cross-validation folds (other folds differ by at most a few objects); the test set is fixed and shared by all folds. Unlabeled objects are used only for self-supervised pretraining and contrastive alignment.

Class	Train	Validation	Test	Total
SN Ia	3377	845	1055	5277
SN II	776	194	243	1213
SESN	288	72	90	450
SN II _n	139	34	43	216
SLSN	101	25	32	158
TDE	39	10	12	61
Total labeled	4720	1180	1475	7375
<i>Unlabeled objects (self-supervised pretraining only, per-modality partitions)</i>				
Images	74620	8268	7836	90724
Light curves	54713	6038	5389	66140
Spectra	8950	1010	163	10123
Distinct unlabeled objects	–	–	–	91906

E Modality specific encoders

Light-curve encoder. The g - and r -band observations are represented as a single temporally ordered sequence. Each measurement is projected into the transformer embedding space, while its observation time is incorporated through feature-wise linear modulation (FiLM). A learnable Fourier representation of time generates multiplicative and additive modulation terms, with separate FiLM transformations for the two bands. Valid observations are then merged and ordered chronologically before being processed by a transformer encoder with a learnable class token and positional embeddings.

During MAE pretraining, contiguous temporal regions are masked rather than isolated epochs, reducing the possibility of solving the reconstruction task through local interpolation alone. Masked tokens remain in the sequence with their input values removed, and a linear head reconstructs the corresponding light-curve measurements.

Spectrum encoder. The spectral encoder follows the same transformer–FiLM formulation, with wavelength replacing time as the conditioning coordinate. Each of the 750 continuum-normalised flux values is embedded and modulated using a learnable Fourier representation of wavelength, with $T_{\max} = 9500 \text{ \AA}$. Contiguous wavelength regions are masked during pretraining, and a linear reconstruction head predicts the missing flux values.

Image encoder. The five image scales are treated as channels of a $5 \times 30 \times 30$ input tensor. The image is divided into 6×6 spatial patches, producing 25 tokens, each spanning the same spatial region across all five scales. The image MAE follows an asymmetric encoder–decoder design: 75% of the patches are removed before the transformer encoder, and a lightweight decoder reintroduces learnable mask tokens to reconstruct the missing patches.

F Training

Masked Auto Encoders Each MAE is pretrained independently on the labeled and unlabeled objects available for its modality by minimizing the mean squared reconstruction error over masked positions. Hyperparameter optimization covers the transformer architecture and optimization parameters, using the reconstruction coefficient of determination, R^2 , on masked validation positions as the objective. Final MAEs are trained with AdamW and a cosine learning-rate schedule, with early stopping and checkpoint selection based on the same validation metric.

Contrastive alignment. The multimodal contrastive stage is initialized from the three pretrained MAEs and uses both labeled and unlabeled objects with at least two available modalities. The encoders remain trainable and are jointly fine-tuned with the projection heads. The search covers the projection architecture and optimization parameters, including the initial inverse temperature. Hyperparameters are selected using the validation separation margin, defined as the mean cosine similarity of matched cross-modal representations minus that of unmatched pairs.

Supervised classification. The unimodal and multimodal classifiers are trained independently using the five stratified folds of Section 3.1. Their pretrained encoders remain trainable during supervised optimization. Classification uses a class-weighted cross-entropy loss, and hyperparameter optimization covers the classifier architecture and optimization parameters, including a discriminative learning-rate multiplier for pretrained components.

Modality drop The multimodal classifiers additionally uses modality-drop augmentation: images, spectra, or both are removed during training by zeroing their representations and clearing the corresponding availability flags. This exposes the GMU to incomplete modality configurations and enables the same trained model to operate with three, two, or one modality at inference. Modality dropping is disabled during evaluation.

Inference modality fine-tuning When evaluating cross-modal transfer, we also consider the case where the MM models are fine-tuned for the specific modalities that will be available at inference time. We keep the alignment, fusion and classification weights unfrozen and fine-tune the models with either LC-only or LC+IM before evaluating with the given input modalities.

Table 2: Hyperparameter search spaces used in the different training stages. “Log” denotes log-uniform sampling and “uniform” linear uniform sampling. The decoder depth is optimised only for the image MAE.

Stage	Hyperparameter	Sampling	Range / choices
MAE ($\times 3$)	Embedding dim. $\times$ heads	categorical	$32 \times 2, 64 \times 2, 64 \times 4, 128 \times 2, 128 \times 4, 128 \times 8, 256 \times 4, 256 \times 8$
	Encoder layers	integer	$[2, 6]$
	Decoder layers [†]	integer	$[1, 4]$
	Dropout	uniform	$[0, 0.3]$
	Learning rate	log	$[10^{-5}, 10^{-3}]$
	Weight decay	log	$[10^{-4}, 10^{-1}]$
	Batch size	categorical	32, 64, 128; spectrum MAE: 16, 32, 64
Contrastive	Projection dimension	categorical	128, 256, 512
	Residual-MLP width factor	categorical	1, 2, 4
	Pooling	categorical	cross-attention; cross-attention without residual MLP; mean pooling
	Projection-head dropout	uniform	$[0, 0.3]$
	Initial inverse temperature	uniform	$[5, 30]$
	Learning rate	log	$[10^{-5}, 10^{-3}]$
	Weight decay	log	$[10^{-4}, 10^{-1}]$
	Batch size	categorical	64, 128, 256
Classifier	MLP hidden dimension	categorical	128, 256, 512
	MLP layers	categorical	1, 2
	MLP dropout	uniform	$[0, 0.4]$
	Backbone learning-rate multiplier	log	$[0.01, 1]$
	Learning rate	log	$[10^{-5}, 10^{-3}]$
	Weight decay	log	$[10^{-4}, 10^{-1}]$
	Batch size	categorical	16, 32, 64

[†]The transformer decoder is used only by the image MAE. The light-curve and spectral MAEs reconstruct masked measurements directly from the encoder outputs using a linear prediction head.

Hyperparameter optimization Hyperparameter selection is based exclusively on validation performance. The optimization objectives are the reconstruction R^2 on masked positions for the MAEs, the cross-modal separation margin for contrastive alignment, and macro-averaged F1 for classification. Parameters inherited from an earlier training stage retain the architecture selected at that stage and are not re-sampled downstream.

Hyperparameters are selected using macro-averaged F1 on the validation split of the first fold. The selected configuration is then trained across all five folds and evaluated on the common held-out test set. Final performance is reported using macro F1 to account for the strong class imbalance.

Hyperparameter searches are performed with Ray Tune using an Optuna TPE sampler and an asynchronous successive-halving (ASHA) scheduler with reduction factor 3 and a grace period of 10 epochs. Each study contains 80 trials in one cross-validation fold. Separate studies are performed for the light-curve, image, and spectral MAEs, the contrastive stage, and the unimodal and multimodal classifiers. Parameters inherited from earlier stages retain their selected architecture and are not re-sampled downstream. Table 2 summarizes the hyperparameter spaces explored at each stage of the training pipeline.

G Tabular features of the feature-based baseline

Table 3 summarizes the feature families. Names are given as base feature names: quantities marked “ g, r ” are evaluated independently in the two bands, whereas “joint” denotes a quantity derived from both bands together.

The feature-selection procedure follows the exclusions adopted by the operational ALerCE classifier. Specifically, forced-photometry quantities, which are undefined for the present alert-based sample, are excluded together with `Coordinate_x`, `Coordinate_y`, and `Coordinate_z`. We additionally

Table 3: Feature families used by the random-forest baselines. The *LC-only* model contains 158 light-curve-derived quantities. The *LC+metadata* model contains the same quantities plus 20 contextual features from AllWISE, Pan-STARRS, and ZTF reference-image information.

Family	Features / description	Evaluation	LC-only	LC+catalogue
ZTF colours	<code>g-r_mean</code> , <code>g-r_max</code> , <code>g-r_mean_corr</code> , <code>g-r_max_corr</code>	joint	✓	✓
Statistical descriptors	Amplitude, AndersonDarling, Autocor_length, Beyond1Std, Con, Eta_e, ExcessVar, Gskew, IAR_phi, LinearTrend, MaxSlope, Mean, Meanvariance, MedianAbsDev, MedianBRP, PairSlopeTrend, PercentAmplitude, Pvar, Q31, Rcs, SF_ML_amplitude, SF_ML_gamma, Skew, SmallKurtosis, Std, StetsonK, and related variability statistics	g, r	✓	✓
Harmonic model	Harmonics_mse, Harmonics_chi, Harmonics_mag_1..7, Harmonics_phase_2..7	g, r	✓	✓
MHPS	MHPS_ratio, MHPS_low, MHPS_high, MHPS_non_zero, MHPS_PN_flag, MHPS_ratio_365_30, MHPS_low_365, MHPS_high_30	g, r	✓	✓
Periodicity	Multiband_period, PPE, Power_rate_{1,2,1,3,1,4,2,3,4}, Period_band, and delta_period	mixed	✓	✓
SPM	Parametric supernova-model quantities: SPM_A, SPM_t0, SPM_gamma, SPM_beta, SPM_tau_rise, SPM_tau_fall, SPM_chi	g, r	✓	✓
GP-DRW	GP_DRW_sigma, GP_DRW_tau	g, r	✓	✓
Folded-light-curve statistics	Psi_CS, Psi_eta	g, r	✓	✓
FLEET	Parameters derived from the FLEET transient model retained by the ALerCE feature-selection procedure	g, r	✓	✓
Microlensing	ulens_u0, ulens_tE, ulens_fs, ulens_chi	g, r	✓	✓
TDE	TDE light-curve decay-model parameters retained by the ALerCE feature-selection procedure	g, r	✓	✓
Colour variation	color_variation	joint	✓	✓
Detection statistics	positive_fraction	g, r	✓	✓
Timespan	Total temporal baseline of the light curve	global	✓	✓
AllWISE	W1-W2, W2-W3, W3-W4, and ZTF-WISE colours { <code>g,r</code> }-W{1,2,3,4}	global	—	✓
Pan-STARRS PS1	Pan-STARRS colours and contextual quantities, including <code>ps_g-r</code> , <code>ps_r-i</code> , <code>ps_i-z</code> , and <code>sgscore1</code>	global	—	✓
Reference image	<code>mean_distnr</code> , <code>sigma_distnr</code> , <code>mean_sharpnr</code> , <code>mean_chinr</code>	joint	—	✓
Total			158	178

remove `TDE_mag0`, `fleet_t0`, `fleet_m0`, `ulens_t0`, and `ulens_mag0` in both bands, following the same feature-selection convention used by the ALerCE production classifier in order to avoid explosion time and peak magnitude biases. No additional features are removed in this work.

H Detailed results

Table 4: Macro F1 and expected calibration error (15 bins, raw scores, no post-hoc calibration) on the fixed test set (1475 objects; mean $\pm$ standard deviation over the five folds; see Table 5 for significance), grouped by the modalities available at inference. Multimodal models are evaluated by hiding the missing modalities (zeroed inputs and presence flags off); “FT on” rows are trimodal models fine-tuned on the reduced input. Bold marks, per inference group, the best neural-network macro F1 and the best (lowest) ECE, together with every result whose $\pm 1\sigma$ interval overlaps the best one.

Inference modalities	Training strategy	Macro F1	ECE
LC+IM+SPEC	LC+IM+SPEC (bal.)	0.741 $\pm$ 0.020	0.056 $\pm$ 0.006
	LC+IM+SPEC (imb.)	0.716 $\pm$ 0.026	0.063 $\pm$ 0.008
LC+SPEC	LC+IM+SPEC (bal.)	0.717 $\pm$ 0.022	0.045 $\pm$ 0.012
	LC+IM+SPEC (imb.)	0.688 $\pm$ 0.033	0.054 $\pm$ 0.010
LC+IM	LC+IM+SPEC (imb.)	0.634 $\pm$ 0.014	0.089 $\pm$ 0.012
	LC+IM+SPEC (imb.), FT on LC+IM	0.630 $\pm$ 0.024	0.096 $\pm$ 0.021
	LC+IM+SPEC (bal.)	0.598 $\pm$ 0.021	0.065 $\pm$ 0.019
	LC+IM+SPEC (bal.), FT on LC+IM	0.598 $\pm$ 0.015	0.093 $\pm$ 0.014
	LC+IM	0.578 $\pm$ 0.015	0.081 $\pm$ 0.033
IM+SPEC	LC+IM+SPEC (bal.)	0.644 $\pm$ 0.036	0.066 $\pm$ 0.009
	LC+IM+SPEC (imb.)	0.506 $\pm$ 0.048	0.083 $\pm$ 0.008
SPEC	LC+IM+SPEC (bal.)	0.613 $\pm$ 0.042	0.043 $\pm$ 0.011
	LC+IM+SPEC (imb.)	0.501 $\pm$ 0.059	0.041 $\pm$ 0.016
LC	LC+IM+SPEC (bal.), FT on LC	0.589 $\pm$ 0.016	0.031 $\pm$ 0.008
	LC+IM+SPEC (imb.), FT on LC	0.574 $\pm$ 0.038	0.059 $\pm$ 0.012
	LC+IM+SPEC (imb.)	0.574 $\pm$ 0.015	0.057 $\pm$ 0.011
	LC	0.568 $\pm$ 0.017	0.031 $\pm$ 0.006
	LC+IM	0.507 $\pm$ 0.039	0.073 $\pm$ 0.009
	LC+IM+SPEC (bal.)	0.499 $\pm$ 0.035	0.050 $\pm$ 0.020
IM	LC+IM+SPEC (bal.)	0.263 $\pm$ 0.007	0.145 $\pm$ 0.046
	LC+IM+SPEC (imb.)	0.261 $\pm$ 0.018	0.223 $\pm$ 0.022
	LC+IM	0.231 $\pm$ 0.019	0.153 $\pm$ 0.080
<i>Random forest baselines^a</i>			
LC features + MD	class-balanced weights	0.623 $\pm$ 0.007	0.166 $\pm$ 0.002
LC features	class-balanced weights	0.607 $\pm$ 0.013	0.165 $\pm$ 0.003

^a Random forests trained on tabular features (178 with metadata, MD; 158 without).

Table 5: Cross-modal transfer tests. ΔF_1 is the macro F1 of each model minus that of the model trained only on the inference modalities (LC: 0.568; LC+IM: 0.578), averaged over the five folds. Confidence intervals and one-sided p -values come from a paired bootstrap over the 1,475 test objects (10^4 resamples), each resample shared by both models and all fold models. Bold: significant after Bonferroni correction for the eight tests ($p < 0.00625$). Last column: folds in which $\Delta F_1 > 0$.

Inference	Training strategy	ΔF_1	95% CI	p	folds
LC+IM	LC+IM+SPEC (imb.)	+0.055	[+0.026, +0.087]	< 0.001	5/5
	LC+IM+SPEC (imb.), FT on LC+IM	+0.051	[+0.024, +0.081]	< 0.001	5/5
	LC+IM+SPEC (bal.)	+0.020	[−0.010, +0.052]	0.10	4/5
	LC+IM+SPEC (bal.), FT on LC+IM	+0.020	[−0.007, +0.047]	0.07	4/5
LC	LC+IM+SPEC (bal.), FT on LC	+0.021	[−0.003, +0.045]	0.04	4/5
	LC+IM+SPEC (imb.), FT on LC	+0.006	[−0.023, +0.035]	0.33	3/5
	LC+IM+SPEC (imb.)	+0.006	[−0.024, +0.036]	0.34	3/5
	LC+IM+SPEC (bal.)	−0.069	[−0.110, −0.027]	> 0.99	1/5

Test-set confusion matrices (row-normalized, mean over 5 folds) — training strategy → inference input

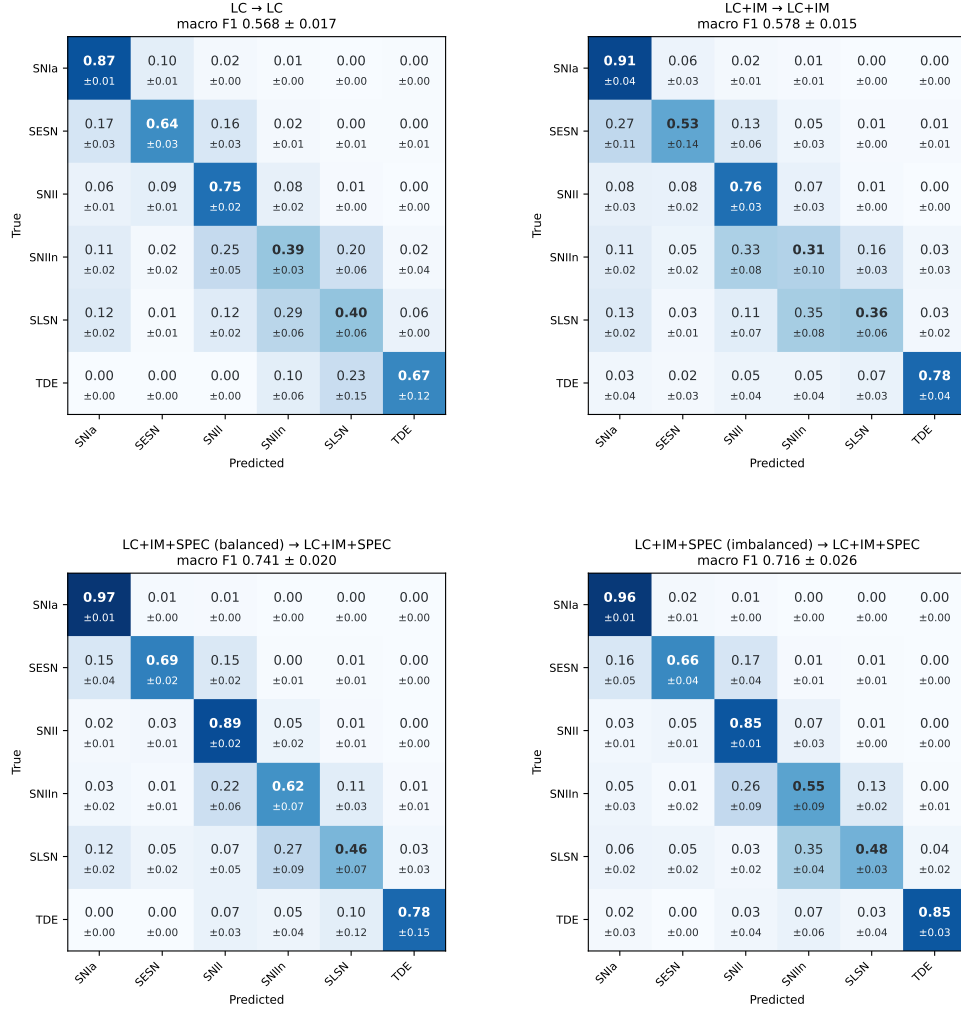


Figure 2: A selection of row normalized confusion matrices evaluated in the test set. We report mean metrics and their standard deviation.

Random forest baselines — test-set confusion matrices (row-normalized, mean over 5 folds)

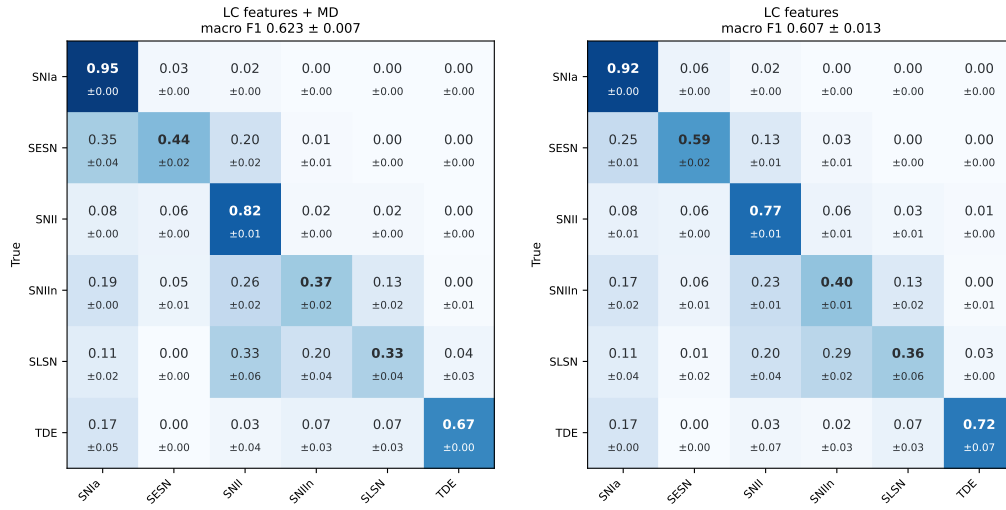


Figure 3: Row normalized confusion matrices evaluated in the test set using the baseline RF model. We report mean metrics and their standard deviation.

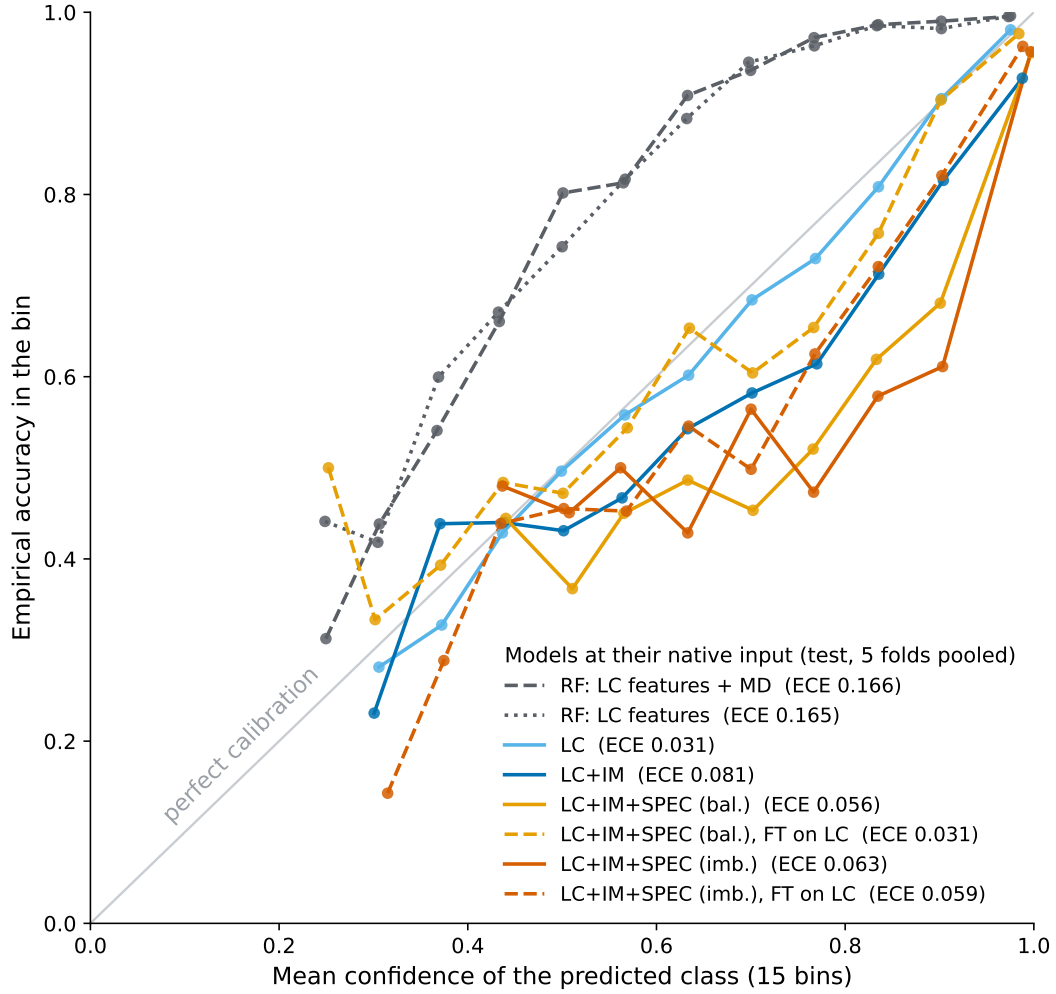


Figure 4: Reliability curves for the RF baseline models and models trained with LCs, LC+IM and LC+IM+SPEC evaluated on their native modalities. Multimodal models have better macro F1 statistics as shown in Table 4, but they tend to be overconfident relative to the LC-only or LC+IM trained models. This effect is reduced when fine-tuning on LC-only. RF baseline models are underconfident, as expected from averaging tree votes; all scores are raw.