A Modernized, AI-Ready Polarized Electron-Positron Dataset

Chi Lung Cheng
SLAC National Accelerator Laboratory
alkaidc@stanford.edu

Simon Corrodi
Argonne National Laboratory
scorrodi@anl.gov

Timothy Hobbs
Argonne National Laboratory
tim@anl.gov

Alaettin Serhan Mete
Argonne National Laboratory
amete@anl.gov

Benjamin Nachman
SLAC National Accelerator Laboratory
nachman@stanford.edu

Abstract

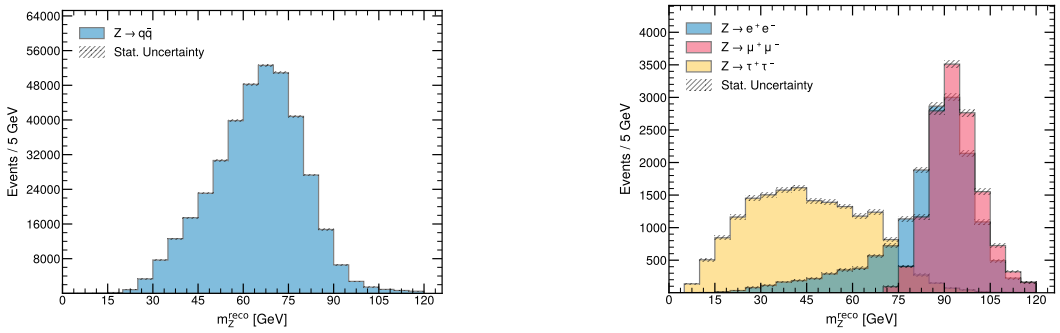
We present an AI-ready public release of reconstructed event data from the SLD experiment at the SLAC Linear Collider, covering the 1996–1998 runs of electron-positron collisions at the Z pole with a longitudinally polarized electron beam. The data have been translated from the legacy Jazelle binary format into modern columnar arrays while preserving the original hierarchical bank structure. As a demonstration of the dataset's capability, we reproduce the historical hadronic and leptonic event selections, validate them with kinematic distributions, and recover values of A_{LR} and A_ℓ consistent with the published SLD results within the limitations of a simplified analysis without Monte Carlo corrections or systematic estimation. To showcase the dataset's value for modern machine learning research in high-energy physics, we examine the SLD jet embedding produced by the Omnilearned foundation model alongside embeddings of other widely used collider datasets, and find that SLD jets occupy a distinct region of the latent space, reflecting the unique experimental topology of polarized e^+e^- collisions that no other publicly available dataset captures.

1 Introduction and Motivation

The shift toward data-driven and machine-learning analysis has reshaped experimental particle physics over the past decade [5], and the diversity of public training and benchmarking datasets is a significant bottleneck. Almost all such datasets come from proton-proton (pp) collisions at the LHC, leaving the e^+e^- and ep regimes underrepresented. Reviving high-quality legacy data broadens the range of physics conditions available to the community for downstream analyses and applications.

The SLD experiment at the SLAC Linear Collider [13, 14] is a uniquely valuable target. Operating from 1992 to 1998, SLD collected over 500,000 Z decays at $\sqrt{s} \approx 91.2$ GeV with a longitudinally polarized electron beam, a configuration not realized at any other collider. This made SLD the source of the most precise direct measurement of the effective weak mixing angle via the left-right asymmetry A_{LR} [2], and SLD remains the only Z-factory dataset where the helicity of the initial state is known event by event.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.



(a) Hadronic $Z \to q\bar{q}$.

(b) Leptonic $Z \to \ell\bar{\ell}$.

Figure 1: Reconstructed visible invariant-mass spectra: (a) hadronic $Z \to q\bar{q}$ channel, (b) the three leptonic Z decay channels overlaid.

The SLD reconstruction software was written in Mortran, a now largely obsolete dialect of Fortran, and the data were stored in the SLD-internal Jazelle binary format, no longer readable by modern analysis tools. The only partial reader we could locate omits the per-event beam polarization required for any polarized observable. We recovered the missing bank layouts by reverse-engineering the binary format with AI assistance (Anthropic's Claude), identifying candidate field positions and data types from recurring byte patterns and alignments across event dumps. Each candidate was validated against physics-based ground truth: the kinematics of well-understood processes such as $Z \to q\bar{q}$ fix the numerical values that must appear in each event. Candidates were also cross-checked bit-for-bit against the partial Java reader on the bank families it covers. Binary decoding errors fail catastrophically rather than subtly (wrong offsets or widths produce unphysical values and break bank-header alignment), so a layout that parses self-consistently across the full sample and reproduces expected kinematics is overwhelmingly likely to be correct. This paper presents the resulting translation of approximately 660,000 reconstructed events from the SLD mini Data Summary Tapes (mini-DSTs), spanning 1996–1998, into a modern columnar format, together with validation studies and a demonstration of the dataset's value for ML research via the Omnilearned foundation model [4].

2 Dataset Provenance and Structure

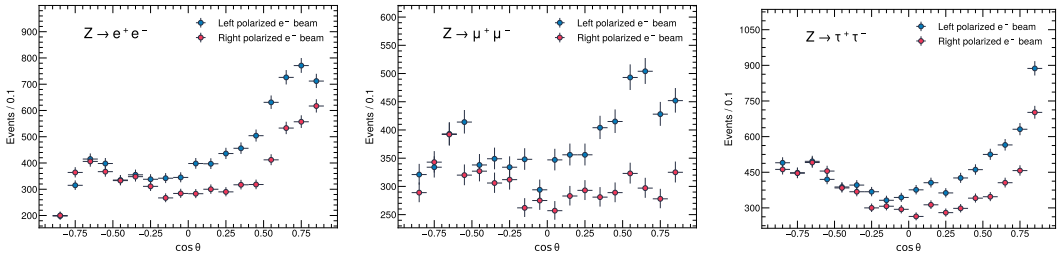
The release covers 1996–1998 SLD operations at the Z pole with a longitudinally polarized electron beam [9]. The 1992–1995 data have not yet been recovered and are not part of this release.

The translation pipeline is released as the open-source Python package `jazelle` [8], which reads the legacy binaries and emits Awkward [11] record arrays that can be serialized to several AI-friendly columnar formats including HDF5, Parquet, and Feather. The original SLD bank hierarchy is preserved. Per-event content covers the major detector subsystems and run conditions, including event header information, beam polarization and its uncertainty, reconstructed physics particles with four-momenta and charges, charged-track parameters from the Central Drift Chamber and VXD3 vertex detector [1], calorimeter clusters from the Liquid Argon Calorimeter, and muon hits from the Warm Iron Calorimeter. Additional mini-DST banks are also preserved.

3 Physics Validation and Measurement Baselines

As a demonstration of the dataset's capability for physics analysis, we reproduce two canonical SLD analyses: the left-right cross-section asymmetry A_{LR} and the leptonic coupling asymmetries A_ℓ.

Event selections. We follow the hadronic event selection of the 2000 A_{LR} analysis [2] and the leptonic selections of the 2001 leptonic-coupling analysis [3], reporting 1996 and 1997–1998 as separate year-groups in line with those analyses.



(a) $\cos\theta$, $Z \to e^+e^-$. (b) $\cos\theta$, $Z \to \mu^+\mu^-$. (c) $\cos\theta$, $Z \to \tau^+\tau^-$.

Figure 2: Polar-angle distributions for the three leptonic channels in the 1997–1998 subset, matching the dataset of the published 2001 leptonic-coupling analysis [3], separated by the sign of the electron-beam polarization.

Table 1: Polarized electroweak observables extracted from the 1996–1998 SLD dataset. $\sin^2\theta_W^{\text{eff}}$ and $\sin^2\theta_W^{\text{lep}}$ are the effective weak mixing angles derived from A_{LR} and from the universality-averaged leptonic asymmetry A_ℓ, respectively. Uncertainties are statistical and beam-polarization only.

	1996	1997–1998	Combined
Hadronic-channel asymmetry			
A_{LR}	0.1554 ± 0.0065	0.1445 ± 0.0041	0.1477 ± 0.0035
$\sin^2\theta_W^{\text{eff}}$	0.2305 ± 0.0008	0.2318 ± 0.0005	0.2314 ± 0.0004
Leptonic-channel asymmetries			
A_e	0.132 ± 0.045	0.165 ± 0.016	0.154 ± 0.012
A_μ	0.175 ± 0.049	0.130 ± 0.018	0.135 ± 0.017
A_τ	0.080 ± 0.046	0.116 ± 0.016	0.112 ± 0.015
A_ℓ (univ.)	—	—	0.137 ± 0.008
$\sin^2\theta_W^{\text{lep}}$	—	—	0.2328 ± 0.0010

Kinematic distributions. Figure 1 shows the reconstructed visible-mass spectra. The hadronic spectrum peaks ~ 20 GeV below M_Z due to energy lost to long-lived neutral hadrons and to neutrinos from heavy-flavor decays. The ee and $\mu\mu$ spectra peak sharply at M_Z as expected for fully visible final states, with a low-energy tail in ee from initial-state radiation, while the $\tau\tau$ spectrum is broad and peaks around 40–50 GeV due to neutrinos in τ decays.

Figure 2 shows the polar-angle distributions of the dilepton final states for the two beam helicities. The differential cross-section for $e^+e^- \to Z \to \ell^+\ell^-$ has a polarization-dependent forward-backward component proportional to $\mathcal{A}_e\mathcal{A}_\ell$, providing the kinematic handle for extracting $\mathcal{A}_\ell$ [3]. Under lepton universality ($\mathcal{A}_e = \mathcal{A}_\ell$), the left- and right-polarized distributions are predicted to cross at $\cos\theta = -1$, a feature clearly visible in the data.

Asymmetry measurements. We extract A_{LR} from the hadronic channel and A_ℓ from the dilepton channels by direct event counting, computed per year-group and combined by inverse-variance weighting. The mean electron-beam polarizations are $\langle\mathcal{P}_e\rangle = (76.14 \pm 1.42)\%$ for 1996 and $(73.08 \pm 1.72)\%$ for 1997–1998. Quoted uncertainties are statistical and beam-polarization only. Results are summarized in Table 1: $\sin^2\theta_W^{\text{eff}} = 0.23144 \pm 0.00044$ from the hadronic channel and $\sin^2\theta_W^{\text{lep}} = 0.23278 \pm 0.00102$ from the leptonic channels.

Scientific caveats. On the hadronic side, our event counts and raw A_{LR} agree with the 2000 high-precision analysis [2] to better than 0.1%. The published analysis, however, converts $A_{LR}(E_{\text{cm}})$ to the pole asymmetry A_{LR}^0 via a small additive correction for γ/Z interference and initial-state radiation computed with ZFITTER 6.22. ZFITTER is no longer runnable on the released sample, so our $\sin^2\theta_W^{\text{eff}}$ differs slightly from the published value by an amount consistent with the omitted correction. Intermediate analysis outputs were not preserved, so event counts are the most upstream comparison available. On the leptonic side, our selected counts differ from those of the 2001 pa-

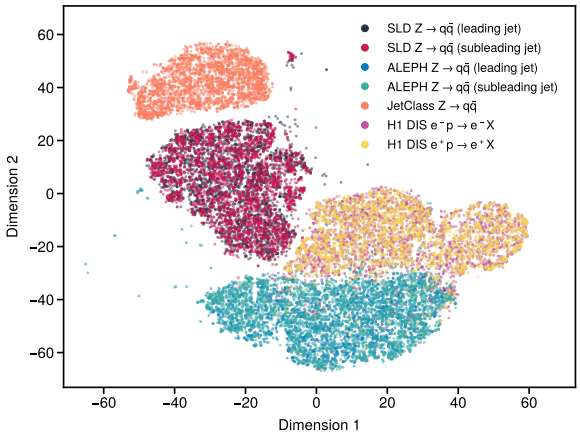


Figure 3: t-SNE projection of Omnilearned jet embeddings across four datasets: SLD (e^+e^-, Z pole, reconstructed detector data), ALEPH (e^+e^-, Z pole) [7], H1 (ep DIS) [10], and JetClass (pp, LHC scale) [12]. The latter three are Monte Carlo simulations. Within each $Z \to q\bar{q}$ dataset the leading and subleading hemispheres collapse onto the same region, while the four datasets themselves form well-separated clusters that organize by initial-state and energy scale.

per [3] by up to $\sim$20%, most visibly in $\mu\mu$ and $\tau\tau$. Such mismatches are typical when reproducing legacy analyses from publications alone, in which the principal selection criteria are documented but internal thresholds, quality flags, and fiducial details enforced by the original code are rarely specified. The dilepton selections are more sensitive to such fine-grained choices than the hadronic one, and we cannot localize the difference further without the original selection software. We also use direct event counting rather than the published unbinned maximum-likelihood fit, with an analytic fiducial correction in place of per-event acceptance corrections. These limitations, together with the omission of non-statistical systematic uncertainties, account for the modest pull of our universal A_ℓ relative to the 2001 published value.

4 Machine Learning Demonstration: Cross-Dataset Jet Embeddings

We embed SLD jets with the Omnilearned foundation model [4] alongside widely used e^+e^-, ep, and pp collider datasets. SLD is the only reconstructed-data sample in the comparison, while the others are Monte Carlo simulations. Because SLD events at the Z pole have a back-to-back two-jet topology, rather than the single narrow jet on which pp-trained models typically operate, we first cluster each event into two hemispheres with the e^+e^- Durham k_T algorithm [6], re-center each on its jet axis, and embed the hemispheres independently.

Projecting the latent vectors into two dimensions via t-SNE [15] (Figure 3) reveals two patterns. Within each $Z \to q\bar{q}$ dataset the leading and subleading hemispheres collapse onto the same region, as expected for two physically equivalent hemispheres. The four datasets, however, resolve into well-separated clusters: SLD and ALEPH share the same physics process but differ in detector and beam energies, while H1 (ep) and JetClass (pp) differ in both initial state and scale. As the only reconstructed-data sample, SLD's cluster represents real detector data from polarized e^+e^- collisions at the Z pole, a regime unavailable from simulation alone. Releasing it adds a genuinely new sample for ML research in HEP, whether for transfer-learning studies, domain-shift benchmarks, or training and evaluation on a regime current models have not seen.

5 Conclusion

We have released the first AI-ready translation of the SLD reconstructed dataset (approximately 660,000 polarized e^+e^- events from the 1996–1998 runs), accompanied by the original binaries and an analysis pipeline that reproduces the published SLD electroweak measurements. A cross-dataset embedding study further shows that SLD jets occupy a distinct region of latent space relative to existing public datasets, underlining the unique experimental territory the data covers.

Code and data availability

The dataset is available on Zenodo at `https://zenodo.org/records/19925960` [9]. The Python translation toolkit `jazelle` is at `https://github.com/HEP-KE/jazelle_reader`, and the analysis code reproducing the measurements presented here is at `https://github.com/HEP-KE/sld-resurrect`.

Acknowledgments

We are grateful to the SLAC accelerator staff that operated SLC and to the SLD Collaboration for collecting and curating this unique dataset and allowing us to resurrect and release it. In particular, we thank Martin Breidenbach, Tony Johnson, and Su Dong for extensive conversations and for loaning us artifacts related to these data, and Yi (Luna) Chen and our colleagues in the electron-positron alliance for useful discussions. The original SLD running was supported by the U.S. Department of Energy (DOE) and the U.S. National Science Foundation; the Istituto Nazionale di Fisica Nucleare of Italy; the Japan–US Cooperative Research Project on High Energy Physics; and, in the United Kingdom, the Science and Engineering Research Council and its successor the Particle Physics and Astronomy Research Council.

This work was supported by the DOE Office of Science under Contracts DE-AC02-06CH11357 and DE-AC02-76SF00515, and was catalyzed and carried out by the Genesis Mission through the HEP pilot project Knowledge Extraction. We are grateful to Katrin Heitmann and our other collaborators on this project.

References

[1] K. Abe et al. Design and performance of the SLD vertex detector: A 307 Mpixel tracking system. *Nucl. Instrum. Meth. A*, 400:287–343, 1997.

[2] Kenji Abe et al. A High precision measurement of the left-right Z boson cross-section asymmetry. *Phys. Rev. Lett.*, 84:5945–5949, 2000.

[3] Koya Abe et al. An improved direct measurement of leptonic coupling asymmetries with polarized Z bosons. *Phys. Rev. Lett.*, 86:1162–1166, 2001.

[4] Wahid Bhimji, Chris Harris, Vinicius Mikuni, and Benjamin Nachman. Foundation model framework for all tasks involving jet physics. *Phys. Rev. D*, 113(3):032020, 2026.

[5] Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Rev. Mod. Phys.*, 91(4):045002, 2019.

[6] S. Catani, Yu. L. Dokshitzer, M. Olsson, G. Turnock, and B. R. Webber. New clustering algorithm for multi-jet cross sections in e^+e^- annihilation. *Phys. Lett. B*, 269:432–438, 1991.

[7] Yi Chen, Lukas Heinrich, Roman Tornqvist, et al. Open data from the ALEPH experiment at the LEP e^+e^- collider. *arXiv preprint*, 2021.

[8] Chi Lung Cheng. `jazelle_reader`: A Python toolkit for translating SLD Jazelle binaries to AI-friendly formats. `https://github.com/HEP-KE/jazelle_reader`, 2026.

[9] Chi Lung Cheng, Benjamin Nachman, Alaettin Serhan Mete, Timothy Hobbs, and Simon Corrodi. SLD reconstructed mini-DSTs from the 1996–1997 SLC runs. Zenodo dataset, 2026. `https://zenodo.org/records/19925960`.

[10] H1 Collaboration. Long-term data preservation and analysis at H1. *EPJ Web Conf.*, 251:02001, 2021.

[11] Jim Pivarski et al. Awkward Array: Manipulating JSON-like data with NumPy-like idioms. Zenodo, 2024.

[12] Huilin Qu, Congqiao Li, and Sitian Qian. Particle Transformer for jet tagging. *Proc. Mach. Learn. Res.*, 162:18281–18292, 2022.

[13] SLC Design Group. SLC Design Handbook. Technical Report SLAC-R-714, SLAC National Accelerator Laboratory, 1984.

[14] SLD Collaboration. SLD Design Report. Technical Report SLAC-R-273, SLAC National Accelerator Laboratory, 1984.

[15] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.