TECHNOLOGY BRIEF

From Optical Illusion to a Realistic Benchmark for Drug-Discovery

KANT employs KANEL technology to support virtual-screening of chemical libraries on Microsoft's Analog Optical Computer digital twin as a defined full-stack hardware benchmark

Sofia Berloff[1], Pavel Koptev[2], Nikita Krainov[2], Konstantin Malkov[2], Alexander Tropsha[3]

August 2026

[1] University of York, York, United Kingdom; [2] ai1 Technologies; [3] UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, USA

> *"Our breakthrough work on an analog optical computer points to new ways to solve complex real-world problems with much greater efficiency."* Satya Nadella [3]

Executive thesis

Photonic computing has produced encouraging results on inference speed and, in particular, on energy per operation. That is why KAN Technologies is evaluating an optical execution layer for high-volume drug discovery and bioinformatics: these are workloads where one trained model is re-executed across very large compound libraries, so cost and energy per score, not one-off training cost, set the economics. The scale of the potential saving is exactly what this study is designed to make testable rather than assert; the figures below are component-level projections, and the matched full-system benchmark that would confirm a product-level advantage is specified in Section 4.

Virtual screening is a workhorse of computational drug discovery where machine learning (ML) models built on training sets comprising 1000's to 100000's entries (i.e., molecules tested for specific bioactivity) are used to infer target bioactivity for up to 100B molecules in virtual screening libraries. A 100B compound library requires 100B model-inferred scores. At that scale, latency, throughput, energy, and cost per scoring run become product constraints. The KANT study asks a commercially relevant but bounded question: can Microsoft's Analog Optical Computer digital twin (AOC-DT) preserve useful ligand-ranking performance on real high-throughput-screening data, and does its solver profile define a credible target for full-stack hardware validation? This study assesses a retrospective application benchmark of a digital twin, not a measurement from a complete optical screening system.

Objective and status. This is a feasibility study, not a product benchmark. It asks whether an AOC-compatible model can execute a standard cheminformatics virtual-screening workflow at accuracy comparable to established digital models, and it demonstrates applicability of the AOC platform to a research domain not covered in the Nature study [1]. The matched, full-stack measurement that would establish a product-level efficiency advantage is not performed here; it is specified in Section 4 as the required next step.

Using three highly imbalanced PubChem datasets (299,000 to 325,000 compounds; 0.53% to 1.40% actives), one four-cell AOC deep-equilibrium model per run is compared with XGBoost on the same repeated random-stratified holdouts. Both models were executed in software on standard digital hardware, the AOC path through Microsoft's calibrated digital twin, so this is an algorithm-level comparison on identical data rather than a comparison of two physical hardware platforms; every optical latency and energy value in this brief is a projection. As an accuracy metric, we employ PPV@128, which is the fraction of true actives among the 128 highest-ranked test compounds. We compare

model performance using three different molecular representations: 103 quantile-scaled LillyMol descriptors, a 256-bit Morgan fingerprint, and Mixed, the 359-feature concatenation of the two. On Morgan fingerprints, the AOC retains 80.0% to 95.6% of XGBoost PPV@128 across the three assays; across all nine assay-by-molecular representation combinations, the range is 67.7% to 95.6%. For completeness, absolute ROC-AUC values are reported in Table 2, but this standard metric is irrelevant in this use case as in practical drug discovery studies, only a small number of top hit molecules is tested experimentally.

Microsoft's 2025 Nature study established the AOC platform: optical matrix-vector multiplication inside a calibrated fixed-point loop, represented by a differentiable digital twin [1]. KANT applies that hardware platform to a workload where repeated inference has clear economic value. The AOC-DT route is molecular features → learned digital input projection → four recurrent AOC cells → digital output projection → ranking score. XGBoost follows a separate route: the same molecular features → gradient-boosted trees → ranking score. XGBoost is an application baseline, not a matched digital implementation: the model families and class-imbalance handling differ, and the Mixed XGBoost setting was not tuned separately. The benchmark therefore measures application-level usefulness; it cannot isolate the optical cell as the cause of an accuracy difference. Those differences are set out in Table 1 and Section 2 [2].

Fingerprint-containing inputs come closer to the XGBoost PPV@128 value than descriptor-only inputs do, for models built across all three assays, and require fewer batch-level fixed-point steps. The solver stops on an integer count for each inference batch. On AID 504466, Morgan fingerprints stop at 9 steps in effectively every batch; LillyMol averages 22.26 steps with a descriptive standard deviation of 3.57 across correlated split-seed runs. The same ordering recurs across all three assays, but it does not establish a direct "native-binary" mechanism because every representation passes through a learned projection, differs in dimension, scale, sparsity, and chemical information, and is trained separately. Seed-score fusion is a tested mitigation for shortlist sensitivity: on AID 504466, the fused Morgan mean slightly exceeds the XGBoost seed-fusion mean and the fused Mixed mean nearly matches it; gaps relative to XGBoost remain on the other assays and for LillyMol. Each additional seed requires another scoring pass unless models are evaluated in parallel, increasing total work and energy [2].

Commercial case and decision gate

- **Scale:** one 100-million-compound library scored across ten target or ensemble passes entails one billion model evaluations.
- **Predictive utility:** on AID 504466, AOC-DT achieves 0.664 PPV@128, about 85 labeled actives in the top 128 held-out compounds; XGBoost achieves 0.694 and about 89 actives. Predictive quality is therefore comparable rather than superior: the win under test is not accuracy but energy and cost per score at library scale.
- **Two candidate hardware test points:** Morgan, the compact 256-bit fingerprint input, which needs the least digital pre-processing before the optical stage; and Mixed, the 359-feature input, which gives the highest AOC accuracy in all three assays while needing about the same number of solver steps as Morgan.
- **Decision gate:** 721 ns and 332 pJ are recurrence-only projections. For scale, the Nature study projects roughly 500 TOPS per watt for a scaled optical core against up to 4.5 TOPS per watt for dense digital arithmetic on conventional platforms [1]; that is an arithmetic-efficiency contrast, not a measured system comparison. The commercial case requires a frozen, auditable rerun and a matched full-stack benchmark reporting predictive quality, shortlist stability, latency, throughput, energy, and cost per score against a declared digital baseline.

1. What Microsoft proved, and why it matters commercially

Microsoft demonstrated that the AOC can execute recurrent AI and optimization computations through a calibrated optical/electronic loop, with a reported loop iteration of approximately 20 ns and close digital-twin correspondence on the published benchmarks [1]. This matters for chemistry because a trained screening model is used to score every compound in a large library. Once predictive quality is adequate, the value of an accelerator depends on sustained scores per second, energy and cost per score, and system integration. The AOC does not create a new chemical capability. Its distinctive claim is physical rather than chemical: the recurrent step executes optically at a reported loop iteration of approximately 20 ns, at a projected arithmetic efficiency of roughly 500 TOPS per watt against up to 4.5 TOPS per watt for dense digital arithmetic [1]. What is under test is therefore energy and latency per repeated inference, not predictive capability that digital models lack.

The KANT study converts that platform proposition into a specific application benchmark. Virtual screening here means retrospective ranking of held-out compounds after model development. The three representations are 103 quantile-scaled LillyMol descriptors, a 256-bit Morgan fingerprint, and Mixed, the 359-feature concatenation of the two. Morgan fingerprints are a compact, widely used representation for molecular similarity assessment and QSAR modeling [7], but the present AOC model does not perform direct bitwise similarity search: every representation first passes through a learned digital projection. The lower Morgan stopping counts are therefore an empirical operating result, not evidence of a binary-specific optical mechanism [2]. Mixed achieves the highest absolute AOC PPV@128 in all three assays and converges near Morgan, so the two inputs define complementary first hardware tests: a leaner input path and a higher-PPV path. Commercially, the study supplies a real workload, measurable quality targets, and a full-system benchmark plan; it does not yet establish a product-level speedup.

Model notation. Ens-4x-AOC-48 is one trained deep-equilibrium model containing four independent 48-variable recurrent AOC cells. Their outputs form a 192-dimensional state read by the digital output projection. Under the published time-multiplexed convention, each fixed-point step requires four sequential optical passes. This is not an average of four separately trained predictors. Ens-4x-AOC-16 uses a different calibrated recurrent map and is not a capacity-only comparison [2].

From Microsoft's AOC platform to KANT's application benchmark

What Microsoft demonstrated	What KANT / Berloff et al. established
A calibrated optical/electronic fixed-point loop for recurrent AI and optimization, with a reported loop iteration of approximately 20 ns.	A three-assay ligand-based virtual-screening benchmark of an AOC-compatible deep-equilibrium model against XGBoost on identical held-out compounds.
The optical recurrence is only one part of a deployable system; digital projection, data movement, control, conversion, and calibration also matter.	A defined recurrence target of about 721 ns and 332 pJ, together with an explicit requirement for a matched end-to-end hardware benchmark.

2. What the KANT benchmark establishes

The completed study covers PubChem AID 504466, AID 485314, and AID 485341, spanning 0.53% to 1.40% active prevalence. Each assay uses five overlapping 80/20 random-stratified holdouts and two or three training seeds. This is a demanding early-recognition setting: only a small fraction of compounds are active, while the model must concentrate actives near the top of the ranking. Because the holdouts overlap and share compounds, means and standard deviations are descriptive within-library comparisons, not independent prospective estimates. AOC-DT and XGBoost are evaluated on the same processed arrays and held-out compounds, but they are not matched architectures: AOC-DT uses assay-dependent positive-class weighting, XGBoost does not, and the Mixed XGBoost setting borrows Morgan hyperparameters. The results therefore compare complete application pipelines rather than isolate the optical cell [2].

On AID 504466, Morgan PPV@128 is 0.664 for AOC-DT versus 0.694 for XGBoost. At 1.363% assay prevalence, this corresponds to about 85 labeled actives among the top 128 held-out compounds, versus 89 for XGBoost; ROC-AUC is 0.897 versus 0.906. XGBoost is therefore the more accurate model here; the practical size of that advantage is about four labeled actives in a 128-compound shortlist, which is small relative to the cost of ranking the full library that produces the shortlist. Across the three assays, Morgan PPV@128 retention is 95.6%, 84.4%, and 80.0%. Within AOC-DT, Mixed gives the highest absolute PPV@128 in each assay and converges near Morgan. This is not predictive parity; it defines two plausible hardware test points rather than one: Morgan for a leaner digital projection and Mixed for higher absolute PPV@128. No separately optimized ensemble design was evaluated [2].

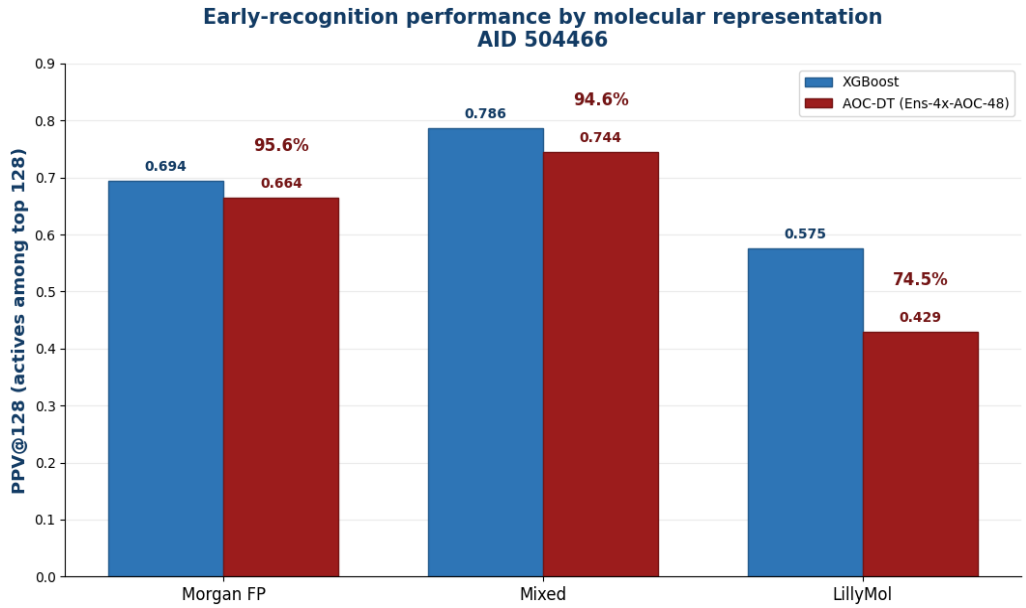


Figure 1: PPV@128 on AID 504466, Ens-4x-AOC-48 versus XGBoost. On Morgan, AOC-DT corresponds to about 85 labeled actives in the top 128 retrospectively held-out compounds versus about 89 for XGBoost. Means over five overlapping holdouts × three seeds.

Early-recognition performance by representation on AID 504466

Representation	XGBoost	AOC-DT (AOC-48)	AOC/XGBoost
Morgan FP	0.694 ± 0.037	**0.664 ± 0.027**	**95.6%**
Mixed (359-dim)	0.786 ± 0.040	0.744 ± 0.047	94.6%
LillyMol descriptors	0.575 ± 0.014	0.429 ± 0.053	74.5%

Table 1: PPV@128 on AID 504466 (mean ± SD over 15 correlated split-seed runs). AOC-DT retains about 95% of XGBoost top-128 precision on Morgan and Mixed inputs; descriptor-only LillyMol shows the largest gap. The representations are not matched for dimension, scale, or information content.

Efficiency status. No matched end-to-end AOC-versus-XGBoost latency or energy ratio is available. The 721 ns and 332 pJ values apply only to the projected optical recurrence. Illustrative dense input-projection scenarios are 21.8 nJ for Morgan and 30.6 nJ for Mixed, before output projection and other overheads. These are scenario calculations, not measurements.

Full ranking-metric comparison on Morgan fingerprints (AID 504466)

Metric (Morgan FP)	AOC-DT (AOC-48)	XGBoost
PPV@128	0.664 ± 0.027	0.694 ± 0.037
PPV@512	0.435 ± 0.015	0.444 ± 0.016
PR-AUC	0.293 ± 0.004	0.306 ± 0.011
ROC-AUC	0.897 ± 0.005	0.906 ± 0.004

Table 2: On Morgan fingerprints, the application-level gap is concentrated at the top 128 and becomes smaller on metrics that integrate more of the ranking. These are descriptive differences, not equivalence tests.

3. Representation-dependent convergence and shortlist stability

Within Ens-4x-AOC-48, the device calibration is shared across representations, but each representation is trained separately and the learned recurrent weights differ. The batch-level stopping condition is reached in consistently fewer iterations for fingerprint-containing inputs: LillyMol requires 2.47 times the Morgan count on AID 504466, 2.27 times on AID 485314, and 1.87 times on AID 485341; Mixed remains close to Morgan [2]. This association recurs across all three assays, but its cause has not been isolated. Raw feature count alone does not explain it: LillyMol has 103 features and the most steps, whereas Mixed has 359 and converges near Morgan's 256-bit input. The data do not establish feature quality as the cause. Every input is projected into a fixed 192-dimensional recurrent state, so input width does not enlarge the recurrent matrix or its per-step optical operation count. Wider inputs can be represented without growing the recurrence, but they may still change the learned recurrence and stopping count. The digital input projection grows linearly: the manuscript's illustrative scenarios are 21.8 nJ for Morgan and 30.6 nJ for Mixed; extrapolating the same arithmetic to 4,096 features gives about 349 nJ before other overheads. These are scenario calculations, not measurements.

The mechanism remains unresolved because the inputs differ in dimension, norm, centering, sparsity, and chemical information, and all pass through a learned digital projection. Ens-4x-AOC-16 and Ens-4x-AOC-48 also use different device calibrations and recurrence structures, so the earlier +9.4/+5.3/+0.6 percentage-point contrast is not a pure capacity test. The second deployment issue is shortlist reproducibility: top-128 shortlists overlap substantially less between AOC training seeds than between XGBoost training seeds [2]. Seed-score fusion raises PPV@128 in several fingerprint-containing settings. On AID 504466, the fused Morgan mean slightly exceeds the XGBoost seed-fusion mean and the fused Mixed mean nearly matches it; residual gaps remain on the other two assays and for LillyMol [2]. This is training-seed sensitivity with a tested mitigation, not a formal bias-variance decomposition. Each additional seed requires another scoring pass unless models are evaluated in parallel; total work and energy increase approximately with seed count, and serial latency does too when passes share one core. One AOC model already contains four recurrent cells, so constituent-model counts are not matched against XGBoost. The next experiment should compare PPV@128, shortlist overlap, latency, and energy at a predeclared fusion level on matched hardware.

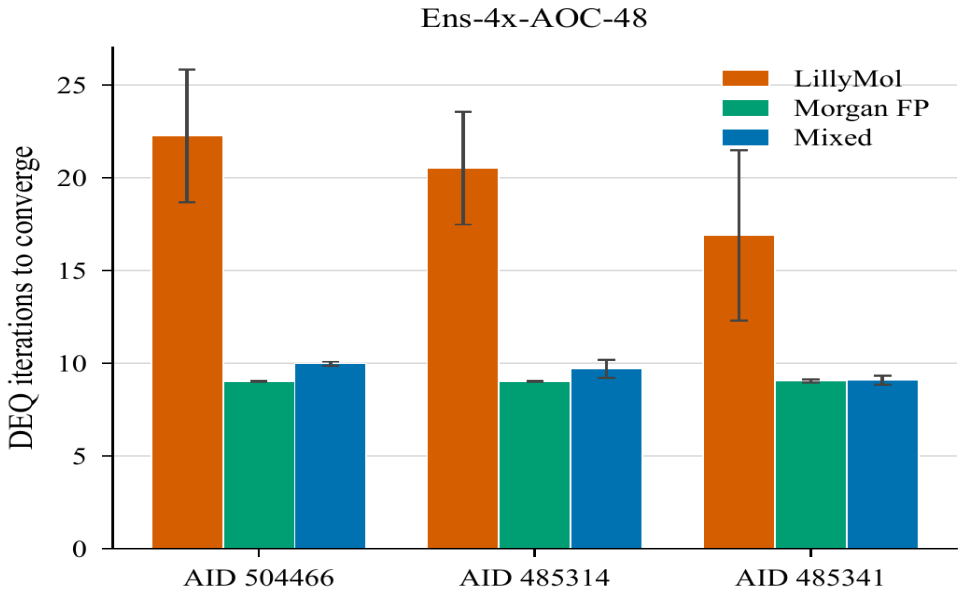


Figure 2: Batch-level fixed-point stopping counts for Ens-4x-AOC-48 across three assays. Morgan remains near 9 steps on all three assays; on AID 504466 it is 9 in effectively every batch. LillyMol requires roughly twice that count, while Mixed converges near Morgan despite containing the full descriptor block. Counts are integer batch-level solver statistics; error bars are descriptive standard deviations across correlated split-seed runs, not per-sample latency distributions.

Assay	LillyMol/Morgan solver-step ratio	Morgan PPV@128 AOC/XGB
AID 504466	2.47×	95.6%
AID 485314	2.27×	84.4%
AID 485341	1.87×	80.0%

Table 3: Lower Morgan stopping counts recur across all three assays while retained PPV@128 varies. Morgan is the leaner fingerprint-containing hardware test; Mixed is a complementary higher-PPV test because it converges near Morgan. Neither result establishes a causal binary-hardware mechanism.

4. Efficiency target and validation boundary

The archived Morgan stopping count for AID 504466 is 9 steps per inference batch, reached in effectively every batch (mean 9.01 across runs, standard deviation 0.03). Under the published convention of four sequential optical passes per step at approximately 20 ns each, this gives a recurrence-only target of about 721 ns and 332 pJ per model inference. These are component projections based on a batch-level stopping count, not measured end-to-end per-compound latency, throughput, or energy. They exclude digital input/output projection, controller, memory traffic, conversion, modulation and detection, calibration, and potentially iterative digital correction. The illustrative dense input-projection scenarios are 21.8 nJ for Morgan and 30.6 nJ for Mixed; the Morgan projection is about 66 times the recurrence figure. These calculations identify digital projection and I/O as the first full-system engineering target, not as a reason to discount the optical core [1,2].

The Nature paper projects approximately 500 TOPS per watt at 8-bit precision for a scaled AOC core and uses up to 4.5 TOPS per watt as a dense digital-arithmetic reference [1]. These are arithmetic-efficiency figures, not an AOC-versus-XGBoost system benchmark. XGBoost tree inference and the hybrid AOC path use different operations and carry different data-movement, conversion, control, calibration, and correction costs. A defensible commercial comparison requires a frozen or predeclared refitted XGBoost baseline, declared hardware, complete-path timing and energy measurement, and fixed PPV@128 and shortlist-stability targets. Power should be reported at a stated sustained

throughput and utilization; it cannot be inferred as a standalone product figure from component energy. Until that benchmark is run, the 721 ns and 332 pJ values are hardware targets, not a system-level advantage.

Validation boundary: what Y-randomization does and does not show

Y-randomization is separate from efficiency: it tests whether enrichment depends on the original feature-label association. On AID 504466 with Morgan fingerprints, full permutation of the development labels produces near-chance ranking. ROC-AUC falls to about 0.50, and PPV@128 falls to 0.005 for the AOC and 0.009 for XGBoost, at or below the 1.363% chance expectation on this assay. At this prevalence, values below the expectation are ordinary sampling variation because a random top-128 list contains only about 1.7 actives on average. The control supports label dependence; it does not exclude every form of structure-level leakage, and it has so far been run on one assay [2].

5. What is de-risked today and what must be proved next

De-risked by the current record	Required for product-level validation
A three-assay retrospective AOC-DT benchmark on 299,000 to 325,000 compounds, using identical held-out compounds for AOC and XGBoost.	Prospective wet-lab enrichment, scaffold generalization, and external-library performance.
On Morgan fingerprints, AOC retains 80.0% to 95.6% of XGBoost PPV@128 across three assays.	Predictive parity, superiority to stronger consensus ensembles, or an agreed quality-loss threshold for deployment.
Fingerprint-containing inputs come closer to XGBoost PPV@128 and show lower batch-level stopping counts in every assay.	A controlled representation ablation and confirmation that the stopping-count pattern transfers to physical hardware.
Full Y-randomization on one assay produces near-chance ranking (Section 4); the recurrence target is 721 ns and 332 pJ (Section 4); seed fusion improves PPV@128 in several fingerprint-containing settings (Section 3).	End-to-end latency, sustained throughput, and energy at fixed predictive-quality and fusion targets; measured physical-hardware-to-digital-twin correspondence.
Fixed split masks and complete saved test predictions support the primary benchmark tables.	A frozen, auditable rerun with complete feature-generation, training, analysis, execution, and intermediate-figure records.

Commercial bottom line

KANT has moved the AOC proposition from a generic photonics promise to a specific, testable drug-discovery workload. The digital twin ranks three real, highly imbalanced screening libraries; on Morgan fingerprints it retains 80.0% to 95.6% of XGBoost PPV@128 across the three assays. On AID 504466, the gap relative to XGBoost corresponds to about four labeled actives in a 128-compound retrospective shortlist. Mixed gives the highest absolute AOC PPV@128 in all three assays while converging near Morgan. This is neither predictive parity nor prospective discovery proof, but it is sufficient to justify a scoped full-stack hardware benchmark.

The opportunity is repeated inference. One 1B virtual screening library requires one billion scores. At that scale, any durable reduction in energy or cost per score is multiplied by the number of scores and therefore compounds directly into total program cost. The study identifies two candidate configurations for a first hardware test. Morgan uses the compact 256-bit fingerprint: it reaches its stopping point in about nine solver steps and needs the least digital pre-processing, so it is the cheaper path to build. Mixed uses the 359-feature input: it gives the highest AOC accuracy and stops in a similar number of steps, but its larger input needs more digital pre-processing. Morgan/ECFP fingerprints are widely used in ligand-based screening [7]. The recurrent state remains 192-dimensional regardless of raw input width, but digital projection cost grows with width and the trained recurrence may change. The commercial case therefore depends on co-designing the digital and optical paths, not on the optical recurrence alone.

The next decisive experiment is a matched end-to-end benchmark. Run the complete AOC inference path on physical hardware and a frozen or predeclared refitted XGBoost baseline on declared standard hardware. Include both Morgan and Mixed, and fix the seed-fusion level in advance. Report PPV@128, shortlist overlap, latency per score, sustained scores per second, joules per score, a declared cost model per million scores, calibration state, digital projections, data movement, conversion, and control. The result would determine whether the projected core efficiency survives system integration. This is a well-scoped engineering benchmark with a clear commercial go/no-go output.

Selected sources

1.) Kalinin, K.P., Gladrow, J., Chu, J. et al., Analog optical computer for AI inference and combinatorial optimization, Nature 645, 354 to 361 (2025). 2.) Berloff, S., Koptev, P., Krainov, N. Malkov, K., Tropsha, A., Representation-Dependent Predictive Performance and Fixed-Point Convergence of an Analog Optical Digital Twin for Ligand-Based Virtual Screening (2026, in preparation). 3.) Satya Nadella, LinkedIn post on Microsoft's analog optical computer (3 September 2025). 4.) Truchon, J.-F., Bayly, C.I., J. Chem. Inf. Model. 47, 488 to 508 (2007). 5.) Bai, S., Kolter, J.Z., Koltun, V., Deep Equilibrium Models, NeurIPS 32 (2019). 6.) Chen, T., Guestrin, C., XGBoost: A Scalable Tree Boosting System, KDD 2016. 7.) Rogers, D., Hahn, M., Extended-Connectivity Fingerprints, J. Chem. Inf. Model. 50, 742 to 754 (2010).

Production note. Numerical benchmark values in this brief are drawn from the manuscript's archived three-assay production grid. Approximate active counts and enrichment factors are direct transformations of PPV@128 and assay prevalence; the library-scale score-count examples (100 million, one billion, and 100 billion) are workload arithmetic; and the 4,096-feature value extrapolates the manuscript's dense-projection formula. The five holdouts overlap, so means and standard deviations are descriptive. No XGBoost booster from the reported grid or complete production execution environment survives; therefore no matched digital latency or energy measurement is reported. The 721 ns and 332 pJ values are recurrence-only projections, while 21.8 nJ for Morgan, 30.6 nJ for Mixed, and 349 nJ for 4,096 features are illustrative input-projection scenarios. The manuscript also records incomplete feature-generation metadata and missing intermediate analysis generators. Product-level validation therefore requires a frozen, auditable rerun. This brief preserves the distinction between retrospective application performance, model architecture, and projected optical hardware.