

Can Public LNP Data Rank the Next Library? A Publication-Cold Screening Audit of LNPDB

George Weale
Columbia University (2025)
george.weale@columbia.edu

Abstract—A useful lipid-nanoparticle (LNP) model should choose experiments in a publication it has never seen. We test that question on 13,982 LNPDB formulations in 37 assay panels from 28 publications. Each outer fold withholds one publication, purges exact test formulations, and performs model selection only inside the remaining panels. Candidate-random nDCG@K is 0.656, but publication-cold nDCG@K is 0.570. A second-cycle budget replay makes the result more specific. Testing 20% of a held-out panel recovers 0.284 of its top decile versus 0.203 under exact random selection, a paired uplift of +0.081. Its pointwise 95% interval is [+0.016, +0.148], while its Bonferroni-7 simultaneous interval is [−0.005, +0.170]. At 50%, uplift is +0.096 with simultaneous interval [−0.026, +0.212]. Panel-level any-hit point estimates are below exact random at both budgets, and their simultaneous intervals include zero. A target-free nearest-publication distance also fails to identify reliable transfers (Spearman +0.216, [−0.196, +0.589]). The moderate-budget point estimates favor the ranking, but the declared seven-budget curve does not establish a budget-specific recall or any-hit uplift. These are retrospective operating characteristics, not prospective discovery evidence.

Index Terms—lipid nanoparticles, virtual screening, grouped validation, data leakage, uncertainty, applicability domain

I. START WITH THE DECISION

Suppose a laboratory has made a new library of 100 LNP formulations. It can test only ten. The practical question is simple: *can a model rank the candidates well enough to put the highest-performing formulations in those ten slots?*

A random train–test split asks an easier question. It mixes candidates from the same paper, chemistry campaign, assay, and normalization procedure across both sides of the split. The model can then use the experimental neighborhood it is supposed to predict. That is useful for interpolation, but it is not evidence about the next study.

This work changes the unit of separation. A whole source publication is hidden until the end of each fold. Exact formulations in that publication are purged from training. Model selection happens inside the remaining data, with complete screening panels held out in turn. The test labels never choose a feature, hyperparameter, uncertainty radius, or support threshold.

The experiment answers three progressively harder questions:

- 1) How much screening performance is visible under a candidate-random split?
- 2) How much survives when the next publication is genuinely unseen?

- 3) Can an unlabeled feature-space diagnostic say when the model should abstain?

The answer to the first question is encouraging. The answers to the second and third are not. That contrast is the main result.

II. DATA BECOME SCREENING PANELS

A. Frozen public source

LNPDB consolidates 19,528 formulations and their experimental context from 42 publications and one commercial source [1]. The frozen CSV used here is commit `fc7c38933b445eb54985014f2b8917606462eec1` of the public LNPDB repository. Its SHA-256 digest is recorded in the run manifest.

The eligibility rule was written before model outcomes were generated. Rows must have the LNPDB endpoint `luminescence_normalized`, a finite numeric value, and a PubMed identifier. A screening panel is one experiment identifier crossed with model type, biological target, administration route, cargo, cargo type, and batching mode. This keeps ranking comparisons inside one declared assay context.

An exact formulation signature contains the ionizable, helper, sterol, PEG, and optional fifth-component structures and ratios, along with buffers, mixing method, dose, and nucleic-acid ratios. Repeated signatures inside a panel are collapsed to their median readout. Panels need at least 30 unique candidates.

Figure 1a shows the resulting audit. Of 19,797 CSV rows, 14,630 use the chosen endpoint, 14,302 also have a numeric value and PubMed identifier, and 13,982 remain after 214 within-panel duplicate rows and small panels are removed. These candidates form 37 panels from 28 publications. LNPDB contains 1,080 exact formulation signatures that occur in more than one eligible publication. This is why row-level randomization is not a safe default.

LNP Atlas is not pooled into the analysis. Its archive is machine-readable, but biological activity is stored in a heterogeneous free-text profile rather than a row-wise normalized luminescence endpoint [2]. Particle size, polydispersity, and encapsulation efficiency are valuable measurements; they answer a different question.

B. One comparable target per panel

Absolute normalized values are not assumed to be comparable across publications. Within each panel, the measured values are converted to percentile ranks from zero to one. The model

learns those ranks from training panels. The held-out panel’s ranks are used only for evaluation.

This transformation narrows the claim. The study asks whether formulation and assay metadata can recover the *ordering within a panel*. It does not claim that a rank of 0.9 in a cell line has the same biological meaning as 0.9 in a mouse tissue.

III. A SPLIT THAT MATCHES THE USE CASE

A. Outer study, inner panel

The outer loop leaves out one PubMed publication. Before fitting, every training row whose exact formulation signature occurs in the test publication is removed. A secondary *lipid-cold* analysis also removes every training row whose ionizable-lipid SMILES occurs in the test publication.

Inside each outer training set, three grouped folds leave complete screening panels out. The inner mean nDCG@K chooses between two ridge models and penalties $\alpha \in \{1, 10, 100\}$:

- *descriptor ridge* uses composition, dose, supplied ionizable-lipid physicochemical descriptors, and hashed assay categories;
- *fragment ridge* adds deterministic character fragments from the supplied SMILES strings.

These are deliberately transparent baselines, not replacements for chemistry-aware graph models such as LiON [3] or experimental platforms such as AGILE [4]. Their purpose is to test whether the pooled record carries robust transferable signal before model complexity becomes the explanation. Grouped and chemistry-aware evaluation is a standard safeguard when molecular neighbors can otherwise cross a split [5], [6].

For comparison, a fixed fragment-ridge model is also evaluated over five candidate-random 80/20 splits. Each panel contributes candidates to both training and test in that comparator. No publication identifier, experiment identifier, formulation identifier, row identifier, target value, panel identifier, or source link is available as a feature in either regime.

B. Top-K utility

For a panel with n candidates,

$$K = \max(3, \lceil 0.1n \rceil). \quad (1)$$

The predicted top-K set is compared with the K highest measured candidates. Because both sets have size K , precision and recall are the same; this paper calls the quantity *top-K hit rate*. Random selection has expected hit rate K/n . Enrichment divides the observed hit rate by that expectation.

Normalized discounted cumulative gain at K (nDCG@K) also rewards placing high-percentile candidates near the top, rather than treating every miss equally. Percentile regret is the difference between the oracle top-K mean percentile and the selected top-K mean percentile. All headline estimates weight panels equally. Pointwise 95% intervals resample whole held-out publications 2,000 times; the budget audit also reports family-wise intervals described below.

C. Uncertainty and abstention

Inner panel-held-out residuals calibrate a symmetric 90% interval for the predicted percentile. This is an empirical transfer interval, not a distribution-free conformal guarantee: the next publication need not be exchangeable with the previous ones [7].

The abstention diagnostic uses no target label. Each panel is represented by the centroid of its numeric, categorical, and string-fragment features. Its support distance is the cosine distance to the nearest training-panel centroid. Inner validation distances freeze the 80th-percentile threshold. A test panel beyond that threshold is rejected. The threshold is not retuned after seeing whether it works.

IV. WHAT SURVIVED THE HARDER SPLIT

A. Random splitting flatters the model

Table I gives the main estimates. The candidate-random comparator reaches nDCG@K 0.656 and hit rate 0.295. Publication-cold evaluation reaches 0.570 and 0.141. Ten of 37 held-out panels recover none of their true top-K candidates.

The publication-cold point estimates exceed analytic random expectation: mean random nDCG@K is 0.521 and mean random hit rate is 0.104. The paired differences are +0.050 nDCG@K and +0.037 hit rate. Their intervals, $[-0.006, 0.110]$ and $[-0.007, 0.093]$, include zero. The evidence is therefore compatible with a modest transferable signal, but it does not establish dependable above-random screening across new publications.

The lipid-cold sensitivity is nearly identical to the primary result. Removing every ionizable lipid seen in the test publication changes nDCG@K by -0.002 and hit rate by -0.006 ; both paired intervals include zero. Exact lipid reuse is not what sustains the already-small average signal. More likely limits are assay heterogeneity, sparse cross-study comparability, and missing protocol variables.

B. The support gate points the wrong way

The prespecified diagnostic also produces a clean negative result. Support distance has Spearman correlation +0.315 with nDCG@K; its 95% interval is $[-0.027, 0.606]$. If distance measured danger, the association should be negative.

The gate accepts 27 of 37 panels. Their mean nDCG@K is 0.544, below the all-panel mean of 0.570. Randomly retaining the same number of panels has mean 0.570 and a 95% range of $[0.538, 0.604]$. The gate therefore does not identify safer transfers. A centroid can look far away because a study explores new but learnable chemistry, or look close while hiding an assay shift. Feature proximity is not a substitute for outcome-relevant validation.

C. Intervals are honest but broad

The 90% cross-panel residual interval attains candidate-weighted coverage 0.894. Its mean width is 0.869 on a percentile scale whose entire width is one. A no-information predictor fixed at 0.5 attains coverage 0.899 with width 0.890. The learned interval is only 0.021 narrower.

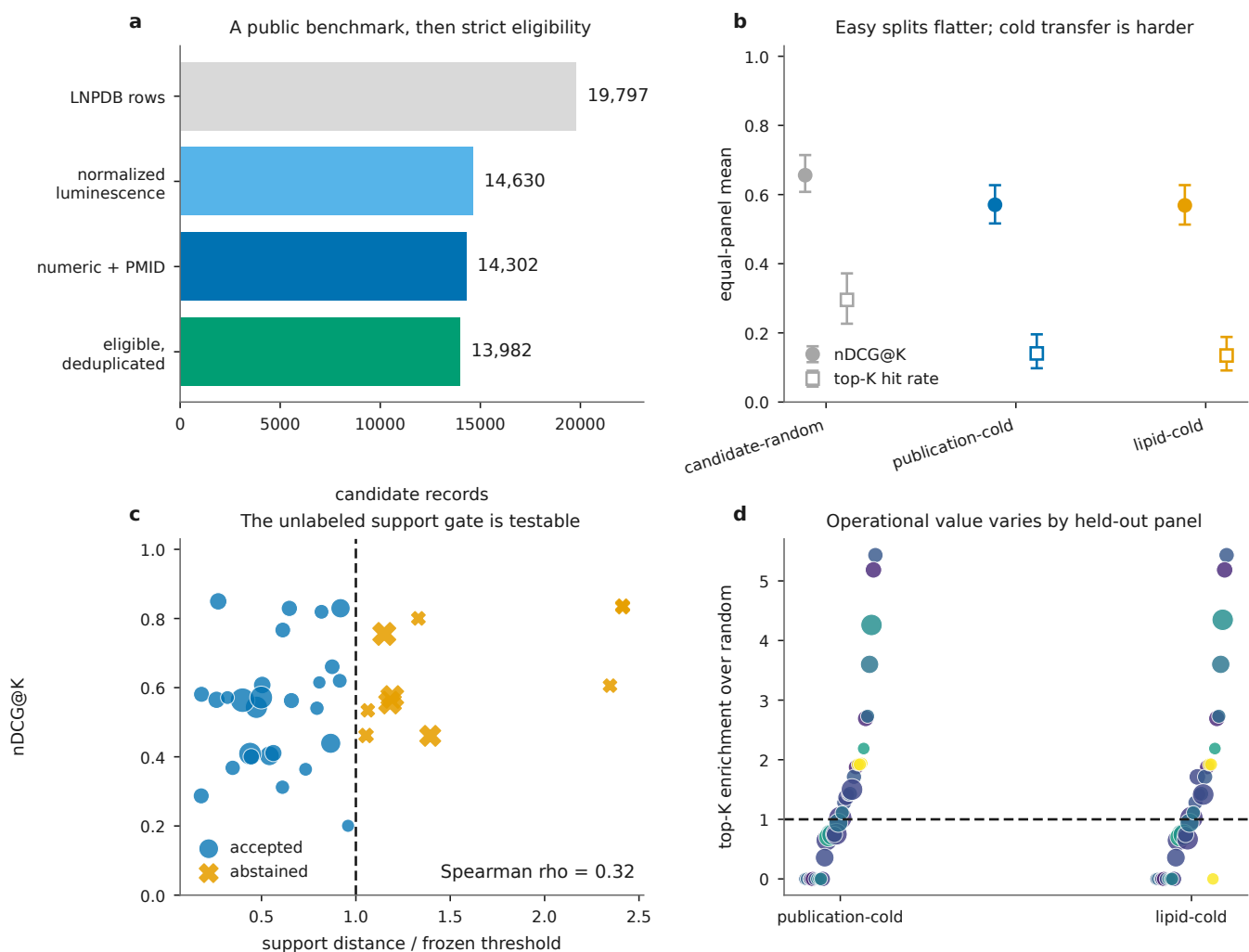


Fig. 1. The complete result in four views. (a) Frozen eligibility audit. (b) Equal-panel means with 95% publication-cluster bootstrap intervals. Candidate-random evaluation is visibly easier than holding out a publication; removing all test ionizable lipids changes little. (c) The target-label-free support gate fails in the prespecified direction: larger distance is not associated with worse nDCG@K, and rejected panels include strong results. Marker area scales with panel size. (d) Enrichment varies sharply by panel; the dashed line is random selection and color encodes support ratio.

This is not a calibration failure. It is an informativeness failure. The interval correctly communicates that a pooled model cannot locate most candidates precisely enough to support a narrow decision.

V. SECOND-CYCLE DECISION AUDIT

Ranking quality is not yet a laboratory decision. A campaign manager must choose a budget, test that prefix, and accept the candidates missed below it. We therefore replayed the frozen publication-cold predictions at seven budgets from 1% to 100%. Random comparators use the exact selected count in each held-out panel, and uncertainty resamples publications as paired clusters.

Recall uplift and any-hit uplift each define a separate family of seven budget comparisons. Their 95% family-wise intervals apply Bonferroni to the publication-cluster bootstrap percentile draws: two-sided limits use quantiles $0.05/(2 \times 7)$ and $1 -$

$0.05/(2 \times 7)$. The two interval sets do not provide joint coverage across all fourteen estimates.

At the 10% nominal budget, the integer rule actually tests 10.37% of candidates. It recovers 0.141 of each panel's top decile, versus 0.104 under exact random selection. Recall uplift is +0.037, with pointwise interval $[-0.006, +0.088]$ and simultaneous interval $[-0.020, +0.109]$. At 20%, recall is 0.284 versus 0.203; uplift is +0.081, with pointwise interval $[+0.016, +0.148]$ and simultaneous interval $[-0.005, +0.170]$. At 50%, recall is 0.597 versus 0.502; uplift is +0.096, with pointwise interval $[+0.008, +0.186]$ and simultaneous interval $[-0.026, +0.212]$.

The easier find-at-least-one question is less favorable. At 20%, 0.811 of panels contain a top-decile hit, below the exact random probability of 0.889. The paired difference is -0.078 , with pointwise interval $[-0.199, +0.028]$ and simultaneous interval $[-0.246, +0.054]$. At 50%, the difference is -0.042 ,

TABLE I
SCREENING PERFORMANCE; EQUAL-PANEL MEANS WITH 95% PUBLICATION-CLUSTER BOOTSTRAP INTERVALS

Evaluation	nDCG@K	Top-K hit rate	Enrichment	Percentile regret
Candidate-random	.656 [.608, .714]	.295 [.227, .372]	1.927 [1.503, 2.430]	.300 [.252, .347]
Publication-cold	.570 [.516, .627]	.141 [.098, .196]	1.357 [.926, 1.863]	.403 [.349, .455]
Publication- and lipid-cold	.569 [.513, .627]	.135 [.091, .188]	1.312 [.890, 1.811]	.404 [.348, .453]

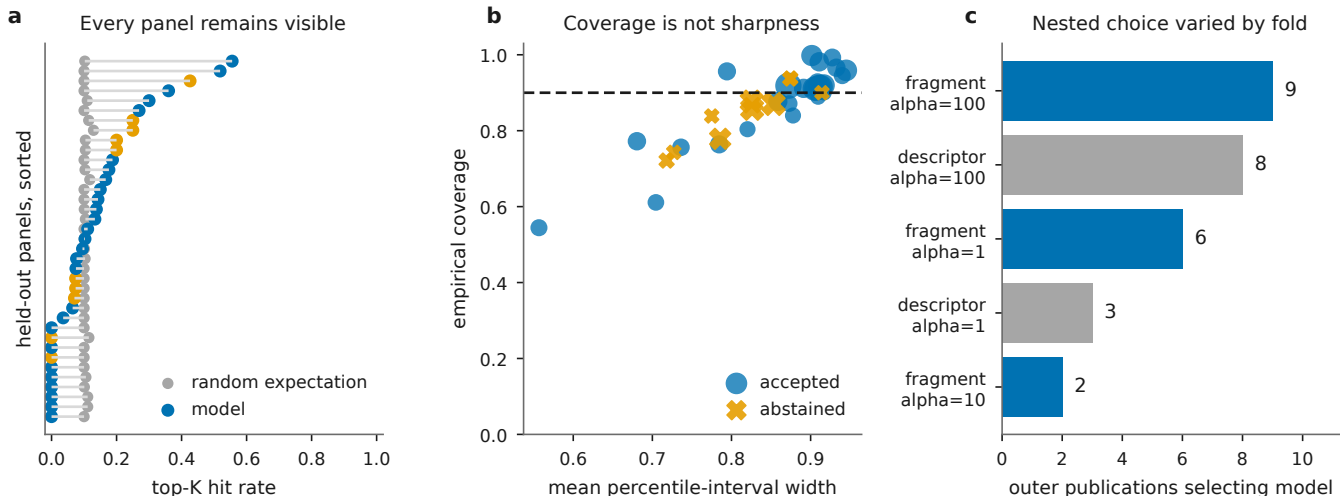


Fig. 2. Failure is not hidden by aggregation. (a) Model and random-expected hit rates for every held-out panel; blue points pass the support gate and ochre points are rejected. (b) Empirical rank-interval coverage versus width. Near-nominal coverage usually requires intervals spanning most of the percentile scale. (c) Nested selection varies across outer folds, so no single model was chosen after viewing all test publications.

with pointwise interval $[-0.109, +0.008]$ and simultaneous interval $[-0.140, +0.010]$. The model can concentrate more good candidates overall without reliably improving the chance that a small campaign finds one.

We also asked whether unlabeled publication similarity could explain transfer. A target-free map compares 754 directed publication pairs using formulation and assay features only. Across 28 held-out publications, nearest-source distance has Spearman correlation $+0.216$ with mean nDCG@K, with interval $[-0.196, +0.589]$. A useful distance warning should become more adverse as distance grows. This map does not.

The moderate-budget recall point estimates exceed random expectation, but their simultaneous intervals include zero. The seven-budget curve therefore does not establish a budget-specific uplift, and the unlabeled support proxy still fails. These are retrospective operating characteristics, not a prospective discovery claim.

VI. WHAT THE RESULT MEANS

The study does not show that machine learning cannot improve LNP design. LiON combines richer molecular representations with prospective experiments [3]; AGILE joins modeling with combinatorial synthesis and assay feedback [4]. Those are stronger evidence chains than retrospective ranking alone.

It does show that an aggregate benchmark can answer the wrong question convincingly. If candidates from one publication

appear on both sides of a split, nDCG@K looks 0.086 higher and hit rate looks 0.155 higher than under the declared new-publication use case. The stricter result is noisy across panels and only marginally above random expectation.

Three changes would make a future public benchmark more decisive:

- 1) publish explicit library and replicate identifiers, failed candidates, assay calibration controls, and measurement uncertainty;
- 2) define a prospective frozen test campaign whose protocols and candidate pool are unavailable during model development;
- 3) evaluate ranking, finite uncertainty, and abstention together, with the abstention rule chosen before the prospective labels exist.

Chemistry-aware fingerprints, graph models, and better support measures remain worth testing. They should face the same publication-cold boundary. A more complex model is progress only if it improves that result without using the held-out publication to choose itself.

VII. LIMITATIONS

LNPDB normalizes heterogeneous source readouts; percentile ranking does not make assays biologically interchangeable. Publication grouping reduces leakage but cannot remove publication bias, extraction error, or unreported protocol

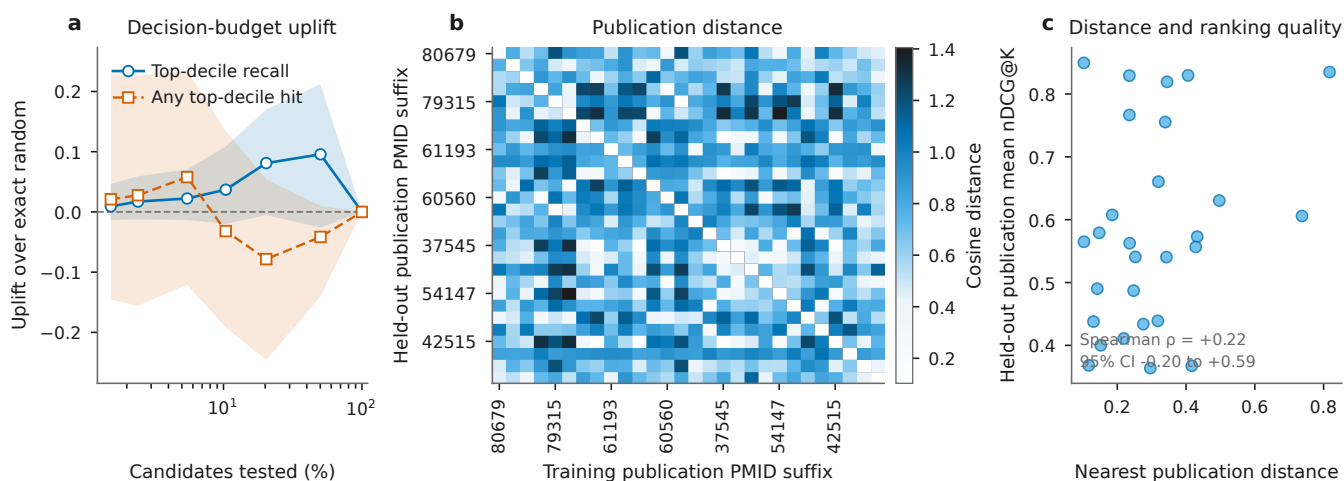


Fig. 3. The second-cycle decision audit. (a) Recall uplift and any-hit uplift over exact random selection as the tested fraction grows. Shaded bands are 95% family-wise Bonferroni-7 intervals, constructed separately for the two outcomes; the horizontal line marks equality with random selection. (b) Target-free cosine distance between held-out and training publications. The diagonal is undefined because a publication is never its own source. (c) Nearest-publication distance does not resolve held-out ranking quality.

differences. The models use supplied descriptors and string fragments, not three-dimensional structure or a learned molecular representation. Panel-centroid distance is only one support diagnostic. Candidate-random and publication-cold evaluation also answer different deployment questions, so their gap is evidence of task difficulty rather than a universal *leakage penalty*.

Most importantly, this is retrospective. It ranks already-published candidates and does not synthesize or test a new LNP. No manufacturing, safety, efficacy, clinical, or causal conclusion follows.

VIII. CONCLUSION

The publication-cold model is not ready to run a campaign. In 13,982 public LNPDB candidates, top-decile recall uplift is +0.081 at 20% and +0.096 at 50%, but the Bonferroni-7 simultaneous intervals $[-0.005, +0.170]$ and $[-0.026, +0.212]$ include zero. At the same budgets, panel-level any-hit point estimates are adverse and their simultaneous intervals include zero; the frozen support gate makes selection worse; nearest-publication distance does not identify reliable transfers; and nominal rank coverage requires almost full-scale intervals.

The moderate-budget point estimates remain useful benchmark targets and prospective hypotheses. They are not a small-budget hit guarantee or abstention claim. Public LNP modeling still needs evaluation units that match real campaigns, richer protocol records, and a prospective frozen test. Until then, a candidate-random score or unlabeled distance should not be presented as evidence that a model can choose the next library’s experiments.

REFERENCES

- [1] E. Collins et al., *Lipid Nanoparticle Database towards structure-function modeling and data-driven design for nucleic acid delivery*, Nature Communications, vol. 17, art. 2464, 2026, doi: 10.1038/s41467-026-68818-1.
- [2] S. Seo, S. Song, and J. Baek, *LNP Atlas: A Comprehensive Database of Lipid Nanoparticle Compositions and Properties for Nucleic Acid Delivery*, Zenodo, 2025, doi: 10.5281/zenodo.17243732.
- [3] J. Witten et al., *Artificial intelligence-guided design of lipid nanoparticles for pulmonary gene therapy*, Nature Biotechnology, vol. 43, pp. 1790–1799, 2025, doi: 10.1038/s41587-024-02490-y.
- [4] Y. Xu et al., *AGILE platform: a deep learning powered approach to accelerate LNP development for mRNA delivery*, Nature Communications, vol. 15, art. 6305, 2024, doi: 10.1038/s41467-024-50619-z.
- [5] Z. Wu et al., *MoleculeNet: a benchmark for molecular machine learning*, Chemical Science, vol. 9, pp. 513–530, 2018, doi: 10.1039/C7SC02664A.
- [6] M. Witman and P. Schindler, *MatFold: systematic insights into materials discovery models’ performance through standardized cross-validation protocols*, Digital Discovery, 2025, doi: 10.1039/D4DD000250D.
- [7] J. Lei, M. G’Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, *Distribution-free predictive inference for regression*, Journal of the American Statistical Association, vol. 113, pp. 1094–1111, 2018.