VERIFIABLE VIRTUAL CELLS: WHEN CAN IN-VITRO DRUG PREDICTIONS BE TRUSTED IN KIDNEY TISSUE?

RESEARCH PLAN • EARLY DRAFT FOR DISCUSSION

Hayat Lab, Uniklinik RWTH Aachen
Draft for internal discussion; author list to be confirmed

4 November 2026

ABSTRACT

Virtual-cell models aim to predict how cells respond to a drug before any experiment, but they are trained almost entirely on cancer cell lines in vitro, while drug discovery needs to know what a drug does in a patient's tissue. For chronic kidney disease (CKD) the gap is basic: in the largest public drug-perturbation atlas, the targets of the main CKD drug classes are essentially not expressed in any cell line, including the only kidney line. Rather than building a better transfer model, we ask when an in-vitro-to-in-vivo prediction can be trusted, and whether this can be judged before the experiment. We reuse the harness of our agent topic: a rule engine checks every prediction on mechanism, data and evaluation code; cases the rules cannot settle go to kidney experts, whose verdicts become new rules. Public in-vivo drug data (CKD mouse and rat studies, one human trial, and tumour positive controls) then measure each prediction's real error, and we test whether predictions judged trustworthy are actually more accurate. This document describes an early design; no results are reported yet.

Keywords virtual cells · perturbation prediction · in vitro to in vivo · verification · rule engine · human in the loop · chronic kidney disease

1 Introduction

Motivation. Virtual cells aim to predict drug effects before an experiment, and almost all are trained on cell lines in vitro. The largest public drug dataset, Tahoe-100M, measured 379 drugs at three doses in 50 cancer cell lines, about a hundred million single cells, under a CC0 licence [1], and a foundation model trained on it is public [2]. What drug discovery cares about, however, is what a drug does in a patient's tissue.

Problem. Even in vitro, complex models often do not beat simple methods [3, 4], and common metrics count systematic differences between treated and control cells as model skill [5]. From in vitro to in vivo, results can flip: the same gene knockouts had opposite-sign effects in some cases between in-vivo and ex-vivo haematopoiesis [6], and predicting human responses from mouse largely failed in a general benchmark [7]. For the kidney the problem is more basic. In Tahoe, the targets of SGLT2 inhibitors, ACE inhibitors and mineralocorticoid-receptor antagonists are essentially not expressed in the only kidney line, and SLC5A2 (SGLT2) is at trace levels at most in all 48 lines with expression data (Appendix B). What a model learns about these drugs in vitro is not how they act in the kidney.

Gap. Existing work makes transfer better: GENEVA measures drug responses of tumour-line pools in mice [8], PerturbRx uses drug transitions learned in vitro to predict whether patients respond [9], and IRIS transfers an in-vitro signalling atlas to mouse embryos [10]. To our knowledge, no work judges, before predicting, whether an in-vitro-to-in-vivo single-cell drug prediction can be trusted, and tests that judgement against measured in-vivo data. Large teams compete on data and compute; our advantage is kidney expertise and checks that can be verified.

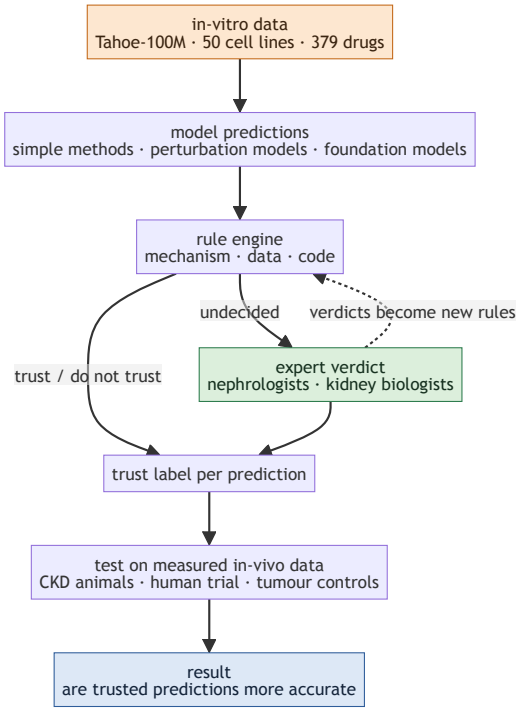


Figure 1: The loop. Models trained in vitro predict for each "drug × in-vivo tissue" pair; the rule engine labels each prediction. Experts (green) see only undecided predictions, and their verdicts become new rules (dotted arrow). The in-vivo data are opened only after labelling, to test whether trusted predictions are more accurate.

Contributions. We plan (i) a trust label for every "drug × in-vivo tissue" prediction, given by a rule engine on three layers (mechanism, data, code); (ii) expert review of the predictions the rules cannot settle, with verdicts turned into versioned rules; (iii) a three-tier design (positive control, partial transfer, mechanism missing) that tests the rules themselves; and (iv) an evaluation of the trust labels against measured in-vivo error in CKD animals and patients. The rule engine and expert process are shared with our agent topic.

2 Approach

Predict, label, then test. Models, simple methods included, learn on Tahoe and predict for every "drug × in-vivo tissue" pair (Figure 1). The rule engine labels each prediction *trust*, *undecided* or *do not trust*, and undecided predictions go to experts. Only then are the in-vivo data opened to measure each prediction's real error. The result we want is not which model is best but whether the labels are right.

Three layers. The *mechanism* layer asks whether the drug has anything to act on in the source cells: the first rule is that a prediction is not trusted when the target is not expressed in the source line; others concern missing tissue context (hormone systems, immune cells, vessels), the gap between in-vitro dose and in-vivo exposure, and whether source and target cells are of the same kind. We draft these rules from the literature and databases, and experts revise them. The *data* layer checks whether the in-vivo data can answer the question: split by animal, with the animal as the statistical unit [11]; scoring that removes systematic variation [5]; and effects that exceed the difference between single-cell and single-nucleus measurement of the same kidneys [12]. The *code* layer checks the evaluation itself: the "no change" and "training mean" baselines must not score perfectly, which happens when test cells leak into the input [13]. Details are in Appendix C.

Three tiers. From the mechanism we know in advance how each tier should behave (Table 1). If the positive control fails, the model or data are at fault; if a model scores well on the mechanism-missing tier, the scoring overrates it.

Experts in the loop. Experts see only predictions the rules cannot settle, mostly in the partial tier, ranked first where rules conflict or the model itself is very uncertain. Nephrologists and kidney biologists judge "trust" or "do not trust"

Table 1: The three tiers. Target expression from DepMap 24Q4 (Appendix B).

Tier	Example drugs (in Tahoe)	Mechanism	Expected
Positive control	KRAS G12C and BRAF V600E inhibitors	target and driver mutation match	transfers
Partial transfer	rosiglitazone (PPARG), dexamethasone (NR3C1)	target present; immune or fat-tissue context missing	experts decide
Mechanism missing	SGLT2 inhibitors, ACE inhibitors, eplerenone	target not expressed in the matched source line	does not transfer

with a one-line reason, which becomes a mechanism-layer rule for the same drug class, with its source recorded (Appendix D). Experts can be wrong, so their verdicts are tested on in-vivo data as well.

3 Evaluation

Trust labels against real error: if the labels are useful, trusted predictions have smaller in-vivo errors than untrusted ones. We compare with three approaches without checks: trust everything, use only the model's own uncertainty, or use only the statistical distance between source and target cells. *Tiers:* the positive control should transfer and the mechanism-missing tier should not. *Kidney and patients:* the mechanism-missing and partial tiers land on CKD in-vivo data, including a human SGLT2-inhibitor trial with dapagliflozin, the same drug as in Tahoe. If simple methods win, nothing changes for this topic: they are among the models being checked (Appendix E).

4 Data and status

The source is Tahoe-100M; its only kidney line is A498, a renal-cancer line with about 2.79 million cells covering all drugs. The in-vivo targets are CKD mouse studies [14, 15], rat studies [16, 17], a human trial with 11 follow-up biopsies [18], and tumour positive controls [8, 19] (Appendix F). The positive control is the weakest part: the three tumour datasets are small and two are mouse. This plan is an early draft: data checks are done, no code exists, and the student, the experts and the timeline are still to be discussed (Appendix G).

References

[1] Jesse Zhang et al. Tahoe-100M: A Giga-Scale Single-Cell Perturbation Atlas for Context-Dependent Gene Function and Cellular Modeling. February 2025. doi: 10.1101/2025.02.20.639398. URL http://dx.doi.org/10.1101/2025.02.20.639398.

[2] Shreshth Gandhi, Farnoosh Javadi, Valentine Svensson, Umair Khan, Matthew G. Jones, John Yu, Daniele Merico, Hani Goodarzi, and Nima Alidoust. Tahoe-x1: Scaling Perturbation-Trained Single-Cell Foundation Models to 3 Billion Parameters. October 2025. doi: 10.1101/2025.10.23.683759. URL http://dx.doi.org/10.1101/2025.10.23.683759.

[3] Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8):1657–1661, August 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02772-6. URL http://dx.doi.org/10.1038/s41592-025-02772-6.

[4] Arc Institute. Virtual Cell Challenge 2025 Wrap-Up. Blog post, 6 December 2025, 2025. URL https://arcinstitute.org/news/virtual-cell-challenge-2025-wrap-up.

[5] Ramon Viñas Torné, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A. Teichmann, Mor Nitzan, and Maria Brbić. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, 44(6):1050–1059, August 2025. ISSN 1546-1696. doi: 10.1038/s41587-025-02777-8. URL http://dx.doi.org/10.1038/s41587-025-02777-8.

[6] Aarathy Ravi Sundar Jose Geetha, Cynthia del Valle, Wolfgang Esser-Skala, David Lara-Astiaso, and Nikolaus Fortelny. Differences between in vivo and ex vivo models drive divergent perturbation outcomes. 2025. doi: 10.1101/2025.09.02.673692. URL http://dx.doi.org/10.1101/2025.09.02.673692.

[7] Jean Radig, Robin Droit, Daria Doncevic, Albert Li, Duc Thien Bui, Luis Herfurth, Thaddeus Kühn, and Carl Herrmann. scArchon: a scalable benchmarking framework for assessing single-cell perturbation models. *Genome Biology*, 27(1), May 2026. ISSN 1474-760X. doi: 10.1186/s13059-026-04104-z. URL http://dx.doi.org/10.1186/s13059-026-04104-z.

[8] Johnny X. Yu et al. The GENEVA platform models tumor mosaicism to reveal variations of responses to KRAS inhibitors and identify improved drug combinations. *Nature Cancer*, 7(3):522–537, February 2026. ISSN 2662-1347. doi: 10.1038/s43018-026-01130-5. URL http://dx.doi.org/10.1038/s43018-026-01130-5.

[9] Yoshitaka Inoue, Minoh Jeong, Alfred Hero, Rui Kuang, and Augustin Luna. PerturbRx: Learning Treatment-Conditioned Latent Transitions for Patient Drug Response Prediction, 2026. URL https://arxiv.org/abs/2608.21349.

[10] Nicholas T. Hutchins, Miram Meziane, Claire Lu, Maisam Mitalipova, David S. Fischer, and Pulin Li. Reconstructing signaling histories of single cells via perturbation screens and transfer learning. *Nature Methods*, 23(10):1973–1983, 2026. ISSN 1548-7105. doi: 10.1038/s41592-026-03213-8. URL http://dx.doi.org/10.1038/s41592-026-03213-8.

[11] Jordan W. Squair et al. Confronting false discoveries in single-cell differential expression. *Nature Communications*, 12(1), 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25960-2. URL http://dx.doi.org/10.1038/s41467-021-25960-2.

[12] Svenja Gaedcke et al. Single cell versus single nucleus: transcriptome differences in the murine kidney after ischemia-reperfusion injury. *American Journal of Physiology-Renal Physiology*, 323(2):F171–F181, August 2022. ISSN 1522-1466. doi: 10.1152/ajprenal.00453.2021. URL http://dx.doi.org/10.1152/ajprenal.00453.2021.

[13] Manuel Gallone, Shyam Ravidath, and Srihith Madireddy. Auditing Confidence-Calibrated Benchmarking of Single-Cell Perturbation Models. October 2026. doi: 10.64898/2026.10.12.758706. URL http://dx.doi.org/10.64898/2026.10.12.758706.

[14] Haojia Wu, Romer Gonzalez Villalobos, Xiang Yao, Dermot Reilly, Tao Chen, Matthew Rankin, Eugene Myshkin, Matthew D. Breyer, and Benjamin D. Humphreys. Mapping the single-cell transcriptomic response of murine diabetic kidney disease to therapies. *Cell Metabolism*, 34(7):1064–1078.e6, 2022. ISSN 1550-4131. doi: 10.1016/j.cmet.2022.05.010. URL http://dx.doi.org/10.1016/j.cmet.2022.05.010.

[15] Jinshan Wu et al. Kidney single-cell transcriptome profile reveals distinct response of proximal tubule cells to SGLT2i and ARB treatment in diabetic mice. *Molecular Therapy*, 30(4):1741–1753, April 2022. ISSN 1525-0016. doi: 10.1016/j.ymthe.2021.10.013. URL http://dx.doi.org/10.1016/j.ymthe.2021.10.013.

[16] Amin Abedini et al. Single-cell transcriptomics and chromatin accessibility profiling elucidate the kidney-protective mechanism of mineralocorticoid receptor antagonists. *Journal of Clinical Investigation*, 134(1), 2024. ISSN 1558-8238. doi: 10.1172/jci157165. URL http://dx.doi.org/10.1172/JCI157165.

[17] Michael S. Balzer, Jianfu Zhou, Amin Abedini, Ziyuan Ma, Yanjuan Hou, Tejaswini Yadav, Manuel Grundmann, Mira Pavkovic, and Katalin Susztak. Single-Cell Resolution Drug Effects of Renin-Angiotensin-Aldosterone Blockade in ZSF1 Rat Diabetic Kidney Disease. *Journal of the American Society of Nephrology*, 37(7): 1438–1456, 2026. ISSN 1533-3450. doi: 10.1681/asn.0000000995. URL http://dx.doi.org/10.1681/ASN.0000000995.

[18] Petter Bjornstad et al. SGLT2 inhibition modulates metabolic, vascular, and inflammatory molecular markers in the kidney in youth with type 1 diabetes. *Science Translational Medicine*, 18(859), 2026. ISSN 1946-6242. doi: 10.1126/scitranslmed.aee1005. URL http://dx.doi.org/10.1126/scitranslmed.aee1005.

[19] Florian Rambow et al. Toward Minimal Residual Disease-Directed Therapy in Melanoma. *Cell*, 174(4): 843–855.e19, August 2018. ISSN 0092-8674. doi: 10.1016/j.cell.2018.06.025. URL http://dx.doi.org/10.1016/j.cell.2018.06.025.

[20] Lanxiang Li et al. A systematic comparison of single-cell perturbation response prediction models. *Science Advances*, 12(37), 2026. ISSN 2375-2548. doi: 10.1126/sciadv.aed3414. URL http://dx.doi.org/10.1126/sciadv.aed3414.

[21] Yusuf H. Roohani et al. Virtual Cell Challenge: Toward a Turing test for the virtual cell. *Cell*, 188(13):3370–3374, June 2025. ISSN 0092-8674. doi: 10.1016/j.cell.2025.06.008. URL http://dx.doi.org/10.1016/j.cell.2025.06.008.

[22] Gabriel M. Mejia, Henry E. Miller, Francis J. A. Leblanc, Bo Wang, Brendan Swain, and Lucas Paulo de Lima Camillo. Diversity by Design: Addressing Mode Collapse Improves scRNA-seq Perturbation Modeling on Well-Calibrated Metrics, 2025. URL https://arxiv.org/abs/2506.22641.

[23] Paul E. Stevens et al. KDIGO 2024 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. *Kidney International*, 105(4):S117–S314, April 2024. ISSN 0085-2538. doi: 10.1016/j.kint.2023.10.018. URL http://dx.doi.org/10.1016/j.kint.2023.10.018.

[24] Broad DepMap. DepMap 24Q4 Public, 2024. URL https://plus.figshare.com/articles/dataset/DepMap_24Q4_Public/27993248/1.

[25] Abhinav K. Adduri et al. Predicting cellular responses to perturbation across diverse contexts with State. 2025. doi: 10.1101/2025.06.26.661135. URL http://dx.doi.org/10.1101/2025.06.26.661135.

[26] Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0494-8. URL http://dx.doi.org/10.1038/s41592-019-0494-8.

[27] Mohammad Lotfollahi et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6), May 2023. ISSN 1744-4292. doi: 10.15252/msb.202211517. URL http://dx.doi.org/10.15252/msb.202211517.

[28] Leon Hetzel, Simon Böhm, Niki Kilbertus, Stephan Günnemann, Mohammad Lotfollahi, and Fabian Theis. Predicting Cellular Responses to Novel Drug Perturbations at a Single-Cell Resolution, 2022. URL https://arxiv.org/abs/2204.13545.

[29] Charlotte Bunne, Stefan G. Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20(11):1759–1768, 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-01969-x. URL http://dx.doi.org/10.1038/s41592-023-01969-x.

A Background and related work

Simple methods and scoring. Five foundation models and two deep models did not beat deliberately simple linear methods at predicting genetic perturbations [3]. Another comparison of 13 methods on 25 datasets found that performance depends strongly on effect size and metric [20]. For the 2025 Virtual Cell Challenge [21], the organisers concluded that models do not yet consistently beat naive baselines across all metrics [4]. Treated and control cells differ systematically, and common metrics reward guessing this difference [5]; metrics should be weighted by differential genes [22].

CKD drugs. ACE inhibitors and angiotensin-receptor blockers block the renin–angiotensin system, SGLT2 inhibitors were first developed for diabetes and later shown to protect the kidney, and mineralocorticoid-receptor antagonists also slow CKD [23]. SGLT2 is the glucose transporter of proximal-tubule cells; ACE and the angiotensin receptor belong to a hormone system that spans organs; the mineralocorticoid receptor acts mainly in the distal nephron. Cancer lines in culture have none of this context.

Table 2: Closest work and what it leaves open.

Work	What it does	What is missing here
GENEVA [8]	pooled tumour lines in mice, single-cell drug responses	cancer only; makes transfer measurable, does not judge trust
PerturbRx [9]	in-vitro drug transitions as features for patient response	predicts response labels, not post-treatment expression
IRIS [10]	in-vitro signalling atlas transferred to mouse embryos	infers which signals acted; no drug responses
Ravi Sundar Jose Geetha et al. [6]	matched in-vivo and ex-vivo Perturb-seq in haematopoiesis	gene knockouts, not drugs; shows the problem
scArchon [7]	benchmark including mouse-to-human transfer	no mechanism checks
Gallone et al. [13]	audits of perturbation benchmarks	no in-vivo endpoint

B Target expression in Tahoe

We read Tahoe's drug and cell-line metadata on Hugging Face and DepMap 24Q4 expression [24] on 2026-11-04. "Expressed" means $\log_2(\mathrm{TPM}+1) > 1$, a lenient threshold; 48 of Tahoe's 50 lines have RNA data. Every Tahoe drug has three doses, 0.05, 0.5 and 5 μM. Tahoe has no empagliflozin, enalapril, lisinopril, ramipril, losartan, irbesartan, finerenone, spironolactone or amiloride.

Table 3: CKD-relevant drugs in Tahoe and their targets.

Class	Drugs in Tahoe	Target	A498	Lines expressing it (of 48)
SGLT2 inhibitors	dapagliflozin, canagliflozin	SLC5A2	0.86	6, at most 1.86, all trace
ACE inhibitors	captopril, perindopril	ACE	0.15	21 (colorectal, pancreatic, lung), not A498 or the liver lines
ARB	only sacubitril/valsartan	AGTR1	0.53	5
MR antagonists	eplerenone	NR3C2	0.36	8, not A498 or the liver lines
PPARγ agonists	rosiglitazone, pioglitazone	PPARG	4.43	35
Glucocorticoids	dexamethasone and others	NR3C1	5.05	36 (most colorectal lines lack it)

Because ACE and NR3C2 pass the threshold in a few non-kidney lines, the mechanism-missing tier is stated as "target absent in the matched tissue line", not "absent everywhere". In DepMap, 7 Tahoe lines carry BRAF V600E and 6 carry KRAS G12C; vemurafenib and adagrasib are in Tahoe.

C Rules in detail

Table 4: First rules, by layer. All are drafts until experts revise them.

Rule	Layer	Source
target not expressed in the source line → do not trust	mechanism	DepMap; this plan
drug needs tissue context absent in vitro (hormone system, immune cells, vessels, fat tissue) → undecided	mechanism	literature; experts
in-vitro dose far above in-vivo exposure → undecided	mechanism	not yet checked
source and target cells of different kind → undecided	mechanism	experts
split by animal; animal as the statistical unit	data	Squair et al. [11]
remove systematic variation before scoring	data	Viñas Torné et al. [5]
effect larger than the single-cell vs single-nucleus difference	data	Gaedcke et al. [12]
"no change" and "training mean" must not score perfectly	code	Gallone et al. [13]
errors weighted by differential genes; scores also with Cell-Eval	code	Mejia et al. [22], Adduri et al. [25]

The single-cell vs single-nucleus data have one sample per condition and assay: enough for a size comparison, not for statistics. Whether Tahoe's top dose of 5 µM is far above the drugs' free plasma levels in patients is not yet checked; the dose rule is written only after that check.

D Experts in the loop

The partial tier needs experts most: the target is present, but which tissue context the drug depends on in the body, and which kidney cell types that matters for, cannot all be written as rules. The system ranks first the predictions where rules conflict or the model is very uncertain. A verdict such as "rosiglitazone acts on the kidney mainly indirectly; do not trust the in-vitro prediction" becomes a mechanism-layer rule for the same drug class, recording the expert and the reason. We report how trust labels from rules, from experts, and from both relate to the real error, rather than whether experts were right. The process is the one designed in our agent topic; who the experts are and how many predictions each judges is not yet decided.

E Models

Three simple methods come first: predict no change, predict the training-set group mean, and a linear model [3]. Existing models are chosen by whether they handle drug data; GEARS predicts only genetic perturbations and is not used. The exact list is set once the simple methods run. Licences were checked on 2026-11-04.

Table 5: Candidate models.

Model	Why	Code licence
scGen [26]	early perturbation model; tested across species	GPL-3.0
CPA [27]	separates drug, dose, time and species	BSD-3
chemCPA [28]	adds molecular structure; unseen drugs	MIT
CellOT [29]	optimal transport; tested on unseen patients and across species	BSD-3
STATE [25]	large model trained on over a hundred million perturbed cells	code CC BY-NC-SA 4.0; weights non-commercial
Tahoe-x1 [2]	foundation model trained on Tahoe-100M	Apache-2.0; weights public

F Datasets

Table 6: Data and their role. Raw data are not redistributed.

Data	Role	Main caveat
Tahoe-100M [1]	source (in vitro, CC0)	one kidney line (A498, renal cancer); three doses per drug
Wu et al. [14] (mouse)	mechanism missing (SGLT2i, ACEi) and partial (rosiglitazone)	4–6 per group
Wu et al. [15] (mouse)	mechanism missing, another lab	drug, sex, duration and assay differ
Abedini et al. [16] (rat)	mechanism missing: MR antagonists	finerenone and spironolactone; Tahoe has only eplerenone; 2–5 per arm
Balzer et al. [17] (rat)	mechanism missing: ACE inhibitor	enalapril; Tahoe has captopril and perindopril; 3 treated rats
Bjornstad et al. [18] (human)	mechanism missing, against patients	dapagliflozin as in Tahoe; 11 follow-up biopsies; no participant IDs, group comparison only
Gaedcke et al. [12] (mouse)	data layer: assay floor	one sample per condition
GENEVA xenograft, GEO GSE311813 [8]	positive control: KRAS G12C	NCI-H2122 is the only shared G12C line, about 1,000 treated cells; ARS-1620 not in Tahoe (adagrasib is)
GEO GSE319598 (mouse lung tumours)	positive control: KRAS G12C	sotorasib vs vehicle, 3–4 tumours per arm; mouse orthologs
GEO GSE116237 [19]	positive control: BRAF V600E	dabrafenib + trametinib, both in Tahoe; about 865 cells; pre-treatment control, no concurrent vehicle

GENEVA's in-vitro 3D pooled cultures can serve as an intermediate tier between 2D in vitro and in vivo.

G Open questions

- **Thin positive control.** Three small tumour datasets; find more, or use GENEVA's 3D cultures as an intermediate tier.
- **Wet lab.** Single-cell data from kidney organoids or tubuloids treated with an SGLT2 inhibitor would fill the gap of kidney drug data in vitro and test the rules directly. To be discussed with the lab and collaborators.
- **Experts.** A few nephrologists and kidney biologists are needed; their time is not yet estimated.
- **Competition.** Several groups work on cross-context transfer; our angle (when to trust, experts in the loop) differs, but timing matters.
- **Still to check.** Tahoe's treatment duration and the gap between its doses and in-vivo exposure; mice per time point in GSE116237; whether the human trial data can be downloaded and used.
- **People and venue.** Which student works on it, and whether a datasets-and-evaluation track at a large ML conference fits; both open.