# Conditioning Ingredient–Permeability Prediction on Measured Substrate Context: A Feasibility Study for Textured–Hair Formulation

Bee Rosa Davis
Davis Geometric
bee_davis@alumni.brown.edu

August 3, 2026

**Abstract**

**Background.** Descriptor-based models of percutaneous absorption predict a permeability coefficient from molecular structure alone. The reference relation of Potts and Guy [1992] has two descriptor terms and a constant, and therefore no slot for the substrate the compound crosses, the vehicle it is delivered in, or the protocol under which it was measured: all such variation is pushed into the residual by construction.

**Methods and results.** We ingest 550 *in vitro* human skin permeability measurements from huskinDB [Stepanov et al., 2020] under a stated drop-and-count policy (550 usable, 0 dropped), of which 534 fall inside the declared applicability domain [Potts and Guy, 1992, OECD, 2007]. Holding descriptors and model form fixed and varying only how much recorded context the fit may condition on, leave-one-out RMSE falls from 1.26 log units under the published coefficients to 1.15 under a global refit of the same 534 rows, and to 0.991 when fits are taken within full (body site × skin layer × donor solution) strata. That last comparison is *not* like-for-like: full conditioning retains only 311 of 534 rows, and we cannot exclude that densely populated protocol cells are intrinsically easier to predict. Against a 29-ingredient textured-hair deck, 12 ingredients (41.4 %) are unanswerable — 8 possessing no single defined molecular structure and 4 lying outside the domain box — with refusals concentrated in the polymers and mixtures that carry the category's structural function. A pre-stated hypothesis, that section curvature $K = \mathrm{Var}/\mathrm{range}^2$ ranks prediction error [Davis, 2026c], failed its own dose-response test and is reported as untestable on pooled data rather than as refuted.

**Scope.** Every measurement analysed is human skin. No hair fibre was measured, and no result here supports a product claim.

**Reproducibility.** Every quantity reported in this paper is computed by scripts in the accompanying repository from a public, CC-BY licensed dataset [Stepanov et al., 2020]. No number is quoted from a secondary source without the arithmetic being reproduced here. Negative and null results are reported in the same detail as positive ones.

**Scope limitation.** All permeability measurements analysed here are *in vitro human skin*. This study contains no hair-fiber measurements, and no result in it supports a product performance claim.

---

# 1 Introduction

A formulator choosing between two conditioning agents for a textured-hair product has no quantitative model of the substrate they are formulating for. This is not a gap peculiar to one product category. The published permeability literature is built almost entirely on skin, and its reference relation — the linear free-energy fit of Potts and Guy [1992] — takes molecular weight and an octanol–water partition coefficient as its only inputs. For a fixed molecule that relation returns a single number, and it returns the same number for every substrate, every anatomical site, every vehicle and every person. Hair fibres differ measurably from one another in cross-sectional ellipticity, twist density and cuticle scale density, and those are the quantities a physical account of surface interaction would need. As far as we have been able to establish — we have not run a systematic review and claim no priority — no published permeability model conditions on any of them. The gap is therefore not a missing coefficient. It is a missing input slot.

We cannot close that gap here, because no public hair-fibre permeability corpus exists to close it with. What can be done is to test, on the one substrate for which a pooled public corpus does exist, whether the architecture such a model would require is worth building. That reduces to three questions. Does conditioning a fit on recorded experimental context reduce error when the molecular representation is held exactly fixed? How much of a realistic ingredient deck can a descriptor-based model answer at all, and what should it return for the rest? And can a dispersion statistic computed by the storage layer serve as a usable confidence signal on data of this shape?

The storage model matters to the first question in a specific way. Records are held in GIGI as sections of a fibre bundle: experimental context lives on the base space and the measured value in the fibre, so a fit can be taken over a fibre rather than over the pooled total space. The curvature statistic we borrow from that layer, $K = \mathrm{Var}/\mathrm{range}^2$, together with the coherence construct built on it [Davis, 2026c], comes from a body of Davis-geometric work [Davis, 2026a,b]; we use it here as a finite-sample dispersion statistic and do not claim it is an instance of the differential-geometric invariants developed there.

Our contributions are:

1. A provenance-preserving ingest of huskinDB [Stepanov et al., 2020] under a drop-and-count policy stated before the data were seen, with the resulting zero published rather than omitted, and a quantification of how sparse the experimental-context metadata actually is: donor solution is unrecorded on 238 of 534 in-domain rows, and only 250 rows specify all three conditioning axes.

2. A four-level conditioning experiment in which the descriptor set and model form are identical at every level and only the depth of context conditioning varies, with every dropped row counted, every rank-deficient cell named, and the non-equivalence of the row sets stated at each level rather than left for the reader to infer.

3. A unit-provenance audit of a redistributed Potts–Guy intercept, in which a bare coefficient copied without its unit convention yields predictions off by a factor of about 2.78, with nothing in the value itself to signal the problem.

4. An applicability-domain and refusal analysis over a 29-ingredient textured-hair deck, separating out-of-domain queries (a data problem, which names the missing measurement) from queries on which the model is undefined (a representation problem, which more data cannot fix).

5. A reported negative result: the section-curvature coherence hypothesis fails its own pre-stated dose-response test, together with the structural reason it cannot be tested on pooled multi-laboratory data at all.

6. A panel design with endpoints and decision rules fixed in advance, whose covariates are restricted to measured fibre geometry.

One scope statement governs all of them. Every measurement analysed in this paper is an *in vitro* human skin permeability coefficient. No hair fibre was measured. The conditioning variables are measured properties of the experimental substrate and protocol, and in the proposed extension they would be measured fibre geometry — never demographic category, race, or self-reported hair type. That is a design constraint on the architecture, stated as such, and nothing measured here bears on how any demographic label would compare to a physical measurement as a predictor.

## 2 Data and provenance

### 2.1 Source and licence

All measured permeability data in this study come from huskinDB, the curated database of *in vitro* human skin permeation experiments assembled by Stepanov, Canipa and Wolber [Stepanov et al., 2020], which is distributed under CC-BY. We pulled the full table from the project's public query endpoint on 2026-08-03 and stored the raw payload verbatim before any processing, so that every downstream number can be traced back to an unedited copy of what the source served. For each field we actually consume — compound name, SMILES, $\log K_p$, body site, skin layer, donor solution, diffusion cell type and reference — the column position was confirmed against the site's own table column definitions rather than assumed.

Two properties of this source govern everything that follows. First, huskinDB records permeability through *human skin*, pooled across many laboratories and protocols; it contains no hair-fibre measurement of any kind. Nothing in this paper is a hair measurement, and no result reported here supports a product claim. Second, the database is an aggregation of published experiments, so its experimental-context metadata is as complete as the underlying papers were — which, as we quantify below, is frequently not very complete at all.

### 2.2 Drop-and-count policy

The endpoint served 550 rows. Our ingest applies a single rule, stated before the data were seen: a row is admitted only if it carries a parseable structure and a numeric measured permeability, and any row failing either test is *dropped and counted*, never imputed. We do not fill missing structures, missing endpoints, or missing context with means, modes, or model-derived guesses at any stage of this study.

The policy was live and it did not fire: all 550 served rows carried a usable SMILES string and a numeric $\log K_p$, and all 550 structures parsed, giving 550 usable records and zero drops. We report the zero explicitly rather than omitting the step, because a drop count is only meaningful if it is published when it is zero as well as when it is not.

### 2.3 Descriptors

huskinDB's payload carries the endpoint but not molecular descriptors, so descriptors are computed by us from the served SMILES using RDKit [RDKit, 2026]. This makes the

descriptors *derived* quantities; the only measured quantity in the record is $\log K_p$, reported base-10 in $\mathrm{cm\,s^{-1}}$.

We compute exactly three: molecular weight (`Descriptors.MolWt`), the Crippen octanol–water partition coefficient estimate (`Crippen.MolLogP`), and topological polar surface area (`Descriptors.TPSA`). The first two are not a modelling choice of ours — they occupy exactly the two descriptor slots of the Potts–Guy baseline [Potts and Guy, 1992], molecular weight and octanol–water partitioning, with the one caveat that Potts and Guy fitted on measured $\log K_{ow}$ whereas we supply a computed Crippen estimate of it. Fixing the descriptor set to the baseline's own inputs is what makes the conditioning comparison in this paper a comparison of context and nothing else. TPSA is computed and stored as a third polarity descriptor available to downstream analysis, but it is not consumed by the two-descriptor model used in the conditioning experiment; we state this so that the stored schema is not mistaken for the fitted feature set.

Across the in-domain records, molecular weight spans $18.0\,\mathrm{g\,mol^{-1}}$ to $585\,\mathrm{g\,mol^{-1}}$ (median $170\,\mathrm{g\,mol^{-1}}$), Crippen $\log P$ spans $-2.56$ to 5.66 (median 1.63), TPSA spans 0 to $211\,\mathrm{Å}^2$ (median $37.3\,\mathrm{Å}^2$), and measured $\log K_p$ spans $-10.9$ to $-1.78$ (median $-6.30$).

## 2.4 Applicability domain

We inherit the applicability domain from the baseline model rather than defining one that flatters our own method. The box we apply is molecular weight $18\,\mathrm{g\,mol^{-1}}$ to $750\,\mathrm{g\,mol^{-1}}$ and $\log P$ from $-3$ to 6, following the range over which Potts and Guy fitted their relation [Potts and Guy, 1992]; treating such a box as the region where descriptor-based prediction is defined is the standard applicability-domain discipline for QSAR-type models [OECD, 2007].

Of the 550 usable records, 534 (97.1 %) fall inside the box and 16 fall outside: 15 on $\log P$ and 1 on molecular weight. The 16 out-of-domain records are retained in the database — they are real measurements — but they are excluded from fitting and scoring, and the exclusion is counted here rather than absorbed silently.

The upper molecular-weight bound is the one place this cut is delicate, and we flag it rather than bury it. The single molecular-weight exclusion is digitoxin at $765\,\mathrm{g\,mol^{-1}}$, which sits just above our stated bound; a reader who prefers a marginally wider box should read the in-domain count as 535 rather than 534, with the $\log P$ exclusions unchanged at 15. Every quantitative result in this paper is computed on the 534 in-domain records, with one stated exception: the dose-response power check on section coherence is deliberately run over all 550 usable records without the domain restriction, and is labelled as such where it is reported.

## 2.5 Storage: records as sections of a fibre bundle

The records are held in GIGI, a database engine whose storage model is a fibre bundle over an explicit base space rather than a flat table: the base space is the key space, the fibre is the value space, and a stored record is a section over its key. We use that structure literally here. The base coordinate is the compound identity — the bundle's declared key is the compound field alone — carried with a per-row disambiguating index so that repeated measurements of the same compound under different protocols remain distinct sections rather than collapsing into one another. The fibre carries the derived descriptors (MW, $\log P$, TPSA), the measured $\log K_p$, the experimental-context fields (body site, skin layer, donor solution), and the source reference. The bundle schema types each field explicitly as numeric or categorical, so the context fields are stored as typed categorical levels rather than as opaque strings, and the stratification reported in this paper groups on those typed fields directly. The section-level construct we later apply

**Table 1:** Composition of the OSHUN dataset. Counts are exact and every reduction between rows is reported, in line with the drop-and-count policy.

| Quantity | Count |
|---|---|
| Rows served by huskinDB [Stepanov et al., 2020] | 550 |
| Dropped: missing SMILES | 0 |
| Dropped: missing $\log K_p$ | 0 |
| Dropped: unparseable structure | 0 |
| Usable records ingested | 550 |
| Excluded, $\log P$ outside $[-3, 6]$ | 15 |
| Excluded, MW outside $18\,\mathrm{g\,mol^{-1}}$ to $750\,\mathrm{g\,mol^{-1}}$ | 1 |
| In-domain records used for all analyses | 534 |
| Distinct compound names (in-domain) | 245 |
| Distinct SMILES strings (in-domain) | 245 |
| Distinct source references (in-domain) | 95 |
| Compounds with more than one record | 88 |
| Compounds with a single record | 157 |
| Compounds with at least three records | 38 |
| Largest number of records for one compound | 16 |

across those context axes is MIRADOR's Layer 3 coherence [Davis, 2026c], whose curvature is the same Davis quantity the engine reports per field, $K = \mathrm{Var}/\mathrm{range}^2$, in the sense developed in Davis [2026a].

All 550 usable records are inserted into the bundle at load time (the in-domain restriction is applied at analysis time, not at write time), and the engine reports 550 records held with a global curvature of 0.013 9 and a corresponding confidence of 0.986 at ingest. A later stage rebuilds the same bundle with only the in-domain rows and the numeric fields for the confidence experiment; the receipt quoted here is from the full load. We report these as load receipts — evidence that the write completed — not as evidence about permeability, and not as a claim that the stored geometry is uniformly well behaved. The same receipt reports a per-record curvature mean of 0.667 against a standard deviation of 12.4 and a 2.5 % two-sigma anomaly rate, so the low global figure sits on top of a heavy-tailed per-record distribution.

### 2.6 Experimental context, and how sparse it is

The conditioning experiments in this paper use three context axes carried by the fibre: anatomical body site, skin layer, and donor solution. A fourth, diffusion cell type, is carried in the extracted records and reported here for completeness, but it is not written into the bundle schema and is not used as a conditioning axis. When the source record leaves a field blank we write the sentinel value `unspecified`; we do not guess the missing site, layer, or vehicle. `unspecified` is therefore a genuine level of each categorical field and is carried as its own stratum wherever we condition, which is why it appears as an explicit stratum in the results rather than as a gap.

Table 2 gives the distinct-value count and the missingness of each field over the 534 in-domain records. The sparsity is substantial and unevenly distributed: skin layer is almost always reported, donor solution almost half the time is not.

**Table 2:** Experimental-context fields available on the 534 in-domain records. "Distinct values" includes the `unspecified` sentinel; "named values" excludes it. Missingness is recorded, never imputed.

| Context field | Distinct values | Named values | `unspecified` | % `unspecified` |
|---|---|---|---|---|
| Body site | 16 | 15 | 81 | 15.2 |
| Skin layer | 7 | 6 | 13 | 2.43 |
| Donor solution | 48 | 47 | 238 | 44.6 |
| Diffusion cell | 9 | 8 | 103 | 19.3 |

The named values are also heavily concentrated. Body site is dominated by abdomen (300 records), followed by thigh (57) and breast (35); skin layer by epidermis (227) and epidermis-plus-dermis (222), with stratum corneum and dermis at 33 each; donor solution, among the records that specify one, by PBS (89) and water (44), with the remaining vehicles in the low tens or fewer — succinate solution 20, NaCl solution 15, citrate–phosphate buffer 14. Diffusion cell type splits mainly between glass diffusion cells (147), Franz cells (145), Pyrex cells (64) and flow-through cells (62). The 534 records come from 95 distinct published references, so this concentration reflects the historical habits of the percutaneous absorption literature, not a sampling decision of ours.

Joint completeness across the three conditioning axes is weaker than any single-axis figure suggests. Only 250 of the 534 in-domain records (46.8 %) specify all three of body site, skin layer and donor solution; 243 specify exactly two, 34 specify exactly one, and 7 specify none. The three axes together realise 100 distinct (site, layer, solution) combinations over 534 records, so the fully conditioned strata are small by construction. This is the central practical constraint on the conditioning experiment: the data support asking whether context reduces error, but they thin out quickly as context is added, and the row counts lost to thin strata are reported alongside every conditioned result rather than being quietly excluded.

A related structural fact matters for the coherence analysis: repeated measurement of the same compound across different contexts is the exception, not the rule. Of the 245 distinct in-domain compounds, 157 appear exactly once and 88 appear more than once; 75 appear under more than one distinct (site, layer, solution) combination, and only 38 appear at least three times. The coherence analysis is restricted to those 38, because the curvature it uses is identically 0.25 for a two-point section and therefore carries no information below three measurements. Of those 38, 34 span more than one recorded context; the remaining four — ephedrine, scopolamine, histidine and methotrexate — carry all of their records under a single recorded context, so for them the curvature reflects within-context replicate scatter rather than variation across the base, and we read their coherence accordingly.

### 2.7 The ingredient query set

Separately from the measured data, we assemble a small deck of 29 INCI-named cosmetic ingredients typical of a textured-hair formulation, used only as queries against the model. This deck is hand-built, not sampled from any database, and carries no measured permeability. Of the 29 entries, 21 have a defined chemical structure — small molecules, together with two quaternary ammonium salts whose stored descriptors include the counter-ion — for which RDKit descriptors can be computed as above; the remaining 8 are polymers, PEG esters, protein hydrolysates or botanical mixtures with no single defined molecular structure. For those 8 we record the absence of a structure by declaration rather than inventing a representative SMILES, since a descriptor-based permeability model is simply undefined on such an input.

The consequences of that declaration are taken up in the results.

### 2.8 What these data cannot tell us

Three limits follow directly from the provenance above and constrain every claim in this paper.

First, the substrate is human skin, pooled across laboratories. It is not hair, and there is no hair-fibre geometry anywhere in this dataset. The conditioning experiment therefore tests whether error lives in measured substrate context on the one substrate for which pooled public data exist; it does not measure hair permeability, and it cannot be read as doing so.

Second, the endpoint is heterogeneous by construction. Records pooled from 95 references differ in cell type, vehicle, donor handling and reporting convention, and roughly half of them decline to state the vehicle at all. Any residual error we report includes this protocol heterogeneity, and we make no attempt to model it away.

Third, our extracted schema carries no donor demographic fields, and this is deliberate as well as factual. The architecture we are evaluating conditions on *measured geometry* — for hair, quantities such as cross-sectional ellipticity and twist density — and never on demographic category, race, or self-reported hair type. That is a design commitment, not a finding of this study: nothing measured here bears on how any demographic label would compare to fibre geometry as a predictor, and we make no such measurement and report no such comparison. We state the commitment because a demographic label is at best a proxy for a physical quantity rather than the quantity itself, and the architecture is only defined on the quantity. Conditioning on the measurement is both better specified and the only version of this architecture we are willing to describe.

## 3 Baseline model

### 3.1 The published relationship

The reference model for percutaneous absorption is the linear free-energy relationship of Potts and Guy [1992], regressed on a compilation of in vitro human skin permeability measurements made from aqueous donor solutions. It predicts the steady-state permeability coefficient $K_p$ from two molecular descriptors, the octanol–water partition coefficient $\log P$ and the molecular weight MW:

$$\log_{10} K_p \;=\; 0.71\,\log P \;-\; 0.006\,1\,\mathrm{MW} \;-\; 6.3, \qquad K_p \text{ in } \mathrm{cm\,s^{-1}}. \tag{1}$$

Two properties of Eq. (1) matter for everything that follows. First, the intercept carries a unit convention: $-6.3$ is the value for $K_p$ expressed in $\mathrm{cm\,s^{-1}}$. Second, the right-hand side is a function of the molecule and of nothing else.

We adopt the $\mathrm{cm\,s^{-1}}$ form throughout this study, which is the convention in which the pooled measurements of Stepanov et al. [2020] are reported, so that predictions and observations are compared on a common scale without any intervening conversion.

### 3.2 Declared applicability domain

The fit is defined over MW from 18 to 750 Da and $\log P$ from $-3$ to $+6$ [Potts and Guy, 1992]. We treat this box as a hard gate rather than a guideline, consistent with the general position that a QSAR carries a declared applicability domain and that predictions outside it are not supported by the model [OECD, 2007]. Of the 550 ingested permeability records, 534 fall

inside the box and 16 fall outside; the 16 are excluded from every comparison scored against Eq. (1), rather than being scored against a model that does not claim to cover them. One later analysis scores no prediction against the published model at all — the dose-response check on section curvature — and the gate is not applied there; that is the single exception, and it is stated again at the point of use.

On those 534 in-domain records, Eq. (1) applied with its published coefficients gives a root-mean-square error of 1.26 log units and a coefficient of determination of $R^2 = 0.021\,0$. Because the coefficients are held fixed rather than refitted, $R^2$ here is $1 - \mathrm{SS}_{\mathrm{res}}/\mathrm{SS}_{\mathrm{tot}}$ taken against the observed mean, not a squared correlation; the two are not interchangeable for a model evaluated out of the box. This is the level-0 reference point against which every conditioning step in Section 4 is measured. An $R^2$ this close to zero means the published relationship, evaluated out of the box on pooled multi-laboratory data, explains almost none of the observed variance — while still, as an ordering heuristic, being one of the most widely redistributed models in the field.

### 3.3 What the model structurally cannot represent

Eq. (1) has exactly three terms and two of them are molecular descriptors. There is no term for the substrate the compound is crossing, no term for the vehicle it is delivered in, no term for anatomical site or tissue preparation, and no term for any geometric property of a fiber. This is not a shortcoming of the fit; it is the shape of the equation. The consequence is exact rather than approximate: for a fixed molecule, the model returns a single number, and that number is the same for every person, every anatomical site, every donor solution and every preparation protocol. All between-subject and between-protocol variation is, by construction, pushed into the residual.

That residual is not small. The pooled records used here span multiple body sites, tissue layers and donor vehicles (Section 2), and a model with no place to put that information must absorb it as error. The question this paper asks is whether the information is recoverable when the model is given somewhere to put it.

One clarification about what "somewhere to put it" means here. The conditioning variables introduced later are measured properties of the experimental substrate and protocol. Where the argument is extended toward textured hair, the intended conditioning variables are measured fiber geometry — and never demographic category, race or self-reported hair type. The reason is structural rather than political: what Eq. (1) lacks is a physical property of the substrate, so the variable that belongs in it is a measurement of that property. A demographic label could enter only as a proxy for that measurement, and a model keyed to the proxy would be answering a question about group membership in place of the question it was actually asked. This paragraph states a design constraint, not a result: no fiber geometry is measured anywhere in this work, and no number reported here quantifies how any such property is distributed in any population.

### 3.4 A unit-provenance hazard in redistributed coefficients

Because the intercept in Eq. (1) is unit-bearing, it changes when $K_p$ is reported in $\mathrm{cm\,h^{-1}}$, which is the more common unit in regulatory practice. Converting is a one-line calculation: since $K_p[\mathrm{cm\,h^{-1}}] = 3\,600\,K_p[\mathrm{cm\,s^{-1}}]$, taking logarithms shifts the intercept and leaves both descriptor slopes untouched,

$$\log_{10} K_p[\mathrm{cm\,h^{-1}}] = \log_{10} K_p[\mathrm{cm\,s^{-1}}] + \log_{10} 3\,600, \qquad \log_{10} 3\,600 \approx 3.556. \tag{2}$$

The converted intercept is therefore

$$c \;=\; -6.3 + \log_{10} 3\,600 \;\approx\; -2.744, \qquad K_p \text{ in } \mathrm{cm\,h^{-1}}. \tag{3}$$

A QMRF-style model card redistributing the same model prints the $\mathrm{cm\,h^{-1}}$ intercept as $-2.3$. The gap between the two values is

$$\Delta \;=\; (-2.3) - (-6.3 + \log_{10} 3\,600) \;\approx\; 0.444 \text{ log units}, \qquad 10^{\Delta} \;\approx\; 2.78, \tag{4}$$

so a prediction made with the printed intercept is a factor of about 2.78 higher than one made with the arithmetic conversion of the published value. Table 3 collects the three numbers.

**Table 3:** The Potts–Guy intercept under three provenance paths. The unit conversion of Eq. (2) is additive and touches the constant only, so the descriptor slopes (0.71 on $\log P$, $-0.006\,1$ on MW) are identical in the first two rows by construction. What slopes accompany the printed value in the third row we have not established.

| Unit convention | Provenance | Intercept |
|---|---|---|
| $\mathrm{cm\,s^{-1}}$ | as published [Potts and Guy, 1992] | $-6.300$ |
| $\mathrm{cm\,h^{-1}}$ | converted, $-6.3 + \log_{10} 3\,600$ | $-2.744$ |
| $\mathrm{cm\,h^{-1}}$ | as printed in a redistributed model card | $-2.300$ |

We state plainly what this is and is not. It is not an accusation of error against any author, compiler or registry. A printed value of $-2.3$ may reflect an independent refit, a different rounding path, a different source regression, or a unit convention we have not identified; we have not established which, and the discrepancy is not the interesting part. The interesting part is that *the coefficient alone cannot tell you.* A bare number $-2.3$, copied out of a model card into a spreadsheet or a script, carries no record of the unit it belongs to, the equation it was fitted in, or the transformation chain that produced it, and a downstream consumer who takes it at face value is off by a factor of nearly three in predicted permeability with nothing in the value itself to signal the problem. Redistribution strips provenance, and stripped provenance is silent.

This is the concrete motivation for the storage model described in Section 2. A coefficient is not a scalar; it is a scalar together with its unit, its source equation, its fitted domain and its citation. Storing it as a bare number is a lossy encoding of what it actually is. A record whose value carries that attached context — a section of a bundle rather than a number in a cell — makes the conversion in Eq. (2) checkable at the point of use rather than reconstructable only by someone who thinks to go back to the 1992 paper. The same argument applies, with higher stakes, to the measurements the model is fitted against: a permeability value pooled from many laboratories is only interpretable alongside the protocol that produced it.

### 3.5 Scope

Everything in this section, and everything measured against this baseline in the rest of the paper, concerns pooled public in vitro human *skin* permeability data — excised tissue in diffusion-cell preparations. Eq. (1) was fitted to skin, our evaluation set is skin, and no hair fiber was measured anywhere in this work. The transposition to textured hair care is an argument about what a model must be able to represent, not a demonstration that it does. No number reported here supports a claim about the behaviour of any ingredient in any product on any person's hair.

# 4 Conditioning on experimental context

## 4.1 Design

The experiment in this section is deliberately narrow. We hold the descriptor set and the model form fixed at every level, and vary one thing only: how much of the recorded experimental context the fit is allowed to condition on. Any change in error is therefore attributable to context, not to a richer molecular representation, a larger hypothesis class, or a better optimiser.

The model at every level is the two-descriptor linear form of Potts and Guy [1992],

$$\log_{10} K_p \;=\; a \log P \;+\; b\,\mathrm{MW} \;+\; c, \tag{5}$$

fitted by ordinary least squares through the normal equations. $\log P$ and MW are computed from each record's SMILES with RDKit [RDKit, 2026]; no descriptor is ever imputed. The evaluation set is the 534 in-domain records described in Section 2, i.e. the 550 usable rows restricted to $18 \le \mathrm{MW} \le 750$ and $-3 \le \log P \le 6$, an explicit applicability-domain box in the sense of OECD [2007].

The four levels are:

**Level 0 — no conditioning, no fitting.** The published Potts–Guy coefficients $a = 0.71$, $b = -0.006\,1$, $c = -6.3$ [Potts and Guy, 1992] applied to our rows as-is. Nothing is estimated from this dataset.

**Level 1 — no conditioning, refit.** Equation (5) refitted on this dataset with a single global set of coefficients. This is the honest baseline: it absorbs whatever systematic offset separates our row set from the corpus Potts and Guy worked with, while still knowing nothing about the experiment that produced each row.

**Level 2 — condition on one context field.** The rows are partitioned by a single recorded context field and Equation (5) is fitted separately within each stratum. We report the three protocol fields we stratify on: body site, skin layer, and donor solution (the vehicle the compound was applied in). The source also records a diffusion-cell type, which we do not stratify on; "context" throughout this section therefore means these three fields, not every field the corpus carries.

**Level 3 — condition on the recorded protocol triple.** Strata are the triple (body site × skin layer × donor solution), with a separate fit in each cell. This is the deepest conditioning we apply, not the deepest the source record would in principle allow.

Every fit at levels 1–3 is scored leave-one-out: each row is predicted from a model fitted to that stratum with that row removed. A stratum is fitted only if it holds at least 8 rows; because each leave-one-out fit uses $n - 1$ rows and estimates three parameters, the thinnest admitted cell is fitted on seven rows with four residual degrees of freedom, which is already marginal and shows up in the results below. A stratum is also left unfitted if its design matrix is rank-deficient, which happens when a cell contains only one or two distinct compounds so that $\log P$ and MW cannot be separated. Rows in strata dropped for either reason are *dropped and counted*. They are never pooled back into a global fit, because doing so would silently reintroduce the unconditioned model under the name of the conditioned one.

RMSE is reported in $\log_{10}$ units, which is not an intuitive scale. We therefore also report the fold-error factor

$$F \;=\; 10^{\,\mathrm{RMSE}}, \tag{6}$$

the multiplicative band on $K_p$ corresponding to one RMSE. Read directly: the published coefficients at level 0 give RMSE = 1.26, i.e. a typical prediction is wrong by a factor of about 18.1 in permeability; the fully conditioned level-3 pooled fit gives RMSE = 0.991, a factor of about 9.80. Both are large. The point of this section is the direction and size of the movement between them, not the absolute accuracy of either. All fold factors and percentage changes quoted in this section are computed from unrounded values; the tables report the same quantities rounded to three significant figures, so recomputing a fold factor from a rounded RMSE will not reproduce the printed figure exactly.

### 4.2 Main result

**Table 4:** Pooled error as a function of how much experimental context the fit conditions on. Model form and descriptors are identical at every level. Fold error is $10^{\mathrm{RMSE}}$, computed from the unrounded RMSE. "Dropped" counts rows excluded from that level's pooled statistics, either because they fall in strata below the 8-row minimum or because their stratum's design matrix is rank-deficient; dropped rows are not pooled back into a global fit.

| Level | Conditioned on | Cells | $n$ | Dropped | RMSE | Fold | $R^2$ |
|---|---|---|---|---|---|---|---|
| 0 | nothing (published coefficients) | — | 534 | 0 | 1.26 | 18.1 | 0.021 0 |
| 1 | nothing (refit on this data) | 1 | 534 | 0 | 1.15 | 14.0 | 0.189 |
| 2 | body site | 6 | 499 | 35 | 1.16 | 14.5 | 0.175 |
| 2 | skin layer | 5 | 528 | 6 | 1.15 | 14.0 | 0.181 |
| 2 | donor solution | 8 | 441 | 93 | 1.10 | 12.7 | 0.274 |
| 3 | site × layer × vehicle | 15 | 311 | 223 | 0.991 | 9.80 | 0.373 |

Table 4 is the central result of the paper. Four things in it are worth stating plainly, including the two that are unflattering.

First, refitting alone buys little. Moving from published coefficients to coefficients estimated on our own 534 rows cuts RMSE by 8.97 % (1.26 to 1.15) and lifts $R^2$ from 0.021 0 to 0.189. The published parameterisation explains about two percent of the variance in this pooled, multi-laboratory row set; a fresh global fit of the same two descriptors explains under a fifth. Molecular structure alone, in this model class, is close to uninformative here.

Second, conditioning on a single context field is mostly a wash. Body site actively hurts: pooled RMSE rises to 1.16 and pooled $R^2$ falls to 0.175, below the level-1 baseline, even though 35 rows in thin site strata were dropped. Skin layer is indistinguishable from baseline (1.15, $R^2 = 0.181$) on essentially the same row set (528 of 534). Only donor solution moves the needle on its own, to 1.10 and $R^2 = 0.274$ — and it does so on 441 rows, having dropped 93.

Third, the protocol triple is where the gain concentrates. Level 3 reaches RMSE = 0.991 and $R^2 = 0.373$: a 13.5 % relative reduction in RMSE against the level-1 refit, and 21.2 % against the published coefficients, with the fold-error band tightening from about 18.1 to about 9.80. No new information about the molecule entered the model to produce this.

Fourth, and this is the limitation that governs how the third point may be read: level 3 covers only part of the data. Of 534 in-domain rows, 329 fall in one of the 17 protocol cells holding at least 8 rows, but only 15 of those cells can actually be fitted, covering 311 rows.

The 223 dropped rows break down as 205 rows in cells below the 8-row minimum and a further 18 rows in the two cells that clear the minimum but contain only two distinct compounds each, leaving their design matrices rank-deficient. The level-3 pooled numbers are therefore *not computed on the same row set* as levels 0 and 1. Some of the apparent improvement may be selection: protocol cells that are densely populated are also cells where one laboratory ran a coherent series, and such rows may be intrinsically easier to predict. We did not measure that confound, and we do not claim to have controlled for it. The level-3 comparison is suggestive of the effect of conditioning; it is not a like-for-like test of it. The same caveat applies in weaker form to level 2 on donor solution (93 rows dropped: 81 in thin strata and 12 in a single rank-deficient stratum of one named vehicle) and in negligible form to level 2 on skin layer (6 rows dropped, all in thin strata).

### 4.3 Per-stratum behaviour

**Table 5:** Level-2 per-stratum leave-one-out results, largest cells within each context field. Six of six fitted body-site cells, five of five fitted skin-layer cells, and the six largest of eight fitted donor-solution cells are shown. Negative $R^2$ means the stratum fit predicts worse than that stratum's own mean.

| Context field | Stratum | $n$ | RMSE | Fold | $R^2$ |
|---|---|---|---|---|---|
| Body site | abdomen | 300 | 1.04 | 10.9 | 0.227 |
| | unspecified | 81 | 1.44 | 27.5 | 0.082 9 |
| | thigh | 57 | 1.25 | 17.6 | 0.137 |
| | breast | 35 | 1.38 | 24.0 | 0.113 |
| | abdomen, breast | 17 | 1.21 | 16.2 | 0.022 9 |
| | abdomen, chest, back | 9 | 0.503 | 3.18 | 0.722 |
| Skin layer | epidermis | 227 | 0.974 | 9.43 | 0.314 |
| | epidermis, dermis | 222 | 1.13 | 13.4 | 0.110 |
| | stratum corneum | 33 | 0.982 | 9.59 | 0.132 |
| | dermis | 33 | 1.63 | 42.3 | −0.313 |
| | unspecified | 13 | 2.46 | 288 | −0.107 |
| Donor solution | unspecified | 238 | 1.34 | 22.0 | 0.179 |
| | PBS | 89 | 0.763 | 5.80 | 0.465 |
| | Water | 44 | 0.849 | 7.06 | 0.199 |
| | Succinate solution | 20 | 0.392 | 2.46 | 0.859 |
| | NaCl solution | 15 | 0.474 | 2.98 | 0.795 |
| | Citrate-phosphate buffer | 14 | 0.814 | 6.51 | 0.367 |

Table 5 explains why the pooled level-2 rows in Table 4 look so flat: the strata are wildly unequal, and what the pooled figure records depends on where the mass sits. In the donor-solution field the largest cell is also the worst-behaved, so it dominates the pooled number. In the body-site and skin-layer fields the largest cell is among the best-behaved — abdomen at RMSE = 1.04, epidermis at 0.974 — and the pooled figure is pulled back up by the smaller, worse cells.

The donor-solution rows make this sharpest. The largest cell is `unspecified` — 238 rows for which the source record does not state the vehicle at all. That cell is not a protocol; it is a record-keeping gap, and it is the worst fitted cell in the field (RMSE = 1.34, fold error

22.0), performing worse than the level-1 global baseline. Every fitted cell in that field with a *named* vehicle beats the baseline, several of them by a wide margin: PBS at 0.763 (fold 5.80), Water at 0.849 (fold 7.06), NaCl solution at 0.474 (fold 2.98), succinate solution at 0.392 (fold 2.46), and the two fitted cells too small to appear in the table, `NaCl, sodium azide` at 0.649 ($n = 13$) and Sorensen buffer at 0.616 ($n = 8$). A ninth named-vehicle stratum, potassium phosphate buffer ($n = 12$), yields no number at all: it holds only two distinct compounds, so its design matrix is rank-deficient and it is dropped. The gain from conditioning is realised on the rows where the context was actually recorded, and it is diluted by the rows where it was not.

Two strata have negative $R^2$: skin layer `dermis` ($n = 33$, $R^2 = -0.313$) and skin layer `unspecified` ($n = 13$, $R^2 = -0.107$). A negative $R^2$ under leave-one-out means the stratum fit predicts held-out rows worse than that stratum's own mean would. Conditioning is not free: splitting the data multiplies the number of estimated parameters and starves each fit, and in thin or internally heterogeneous cells the split costs more than it buys. The `unspecified` skin-layer cell at fold error 288 is the extreme case and should be read as a warning about 13-row fits, not as a finding about dermis or about anything physical.

**Table 6:** Level-3 per-stratum leave-one-out results: the eight largest of 15 fitted full protocol cells. These eight cells hold 236 of the 311 rows covered at level 3; the remaining seven fitted cells hold 75. The 223 rows not covered at level 3 are absent from every number in this table and from the level-3 row of Table 4.

| Body site | Skin layer | Donor solution | $n$ | RMSE | Fold | $R^2$ |
|---|---|---|---|---|---|---|
| abdomen | epidermis | unspecified | 75 | 1.16 | 14.3 | 0.022 9 |
| abdomen | epidermis | PBS | 37 | 0.883 | 7.63 | 0.468 |
| abdomen | epidermis, dermis | unspecified | 28 | 0.925 | 8.42 | −0.120 |
| abdomen | dermis | unspecified | 24 | 1.30 | 20.0 | 0.292 |
| thigh | epidermis, dermis | unspecified | 22 | 1.27 | 18.6 | 0.293 |
| unspecified | epidermis | unspecified | 21 | 0.440 | 2.75 | 0.837 |
| unspecified | epidermis | PBS | 15 | 0.479 | 3.02 | 0.807 |
| breast | epidermis, dermis | Water | 14 | 0.446 | 2.79 | 0.758 |

The level-3 cells in Table 6 show the same split. Cells in which all three context fields are recorded do well — `abdomen / epidermis / PBS` at RMSE = 0.883 ($R^2 = 0.468$), `breast / epidermis, dermis / Water` at 0.446 ($R^2 = 0.758$) — while the largest cell, `abdomen / epidermis / unspecified`, sits at 1.16 with $R^2 = 0.022\,9$, no better than the unconditioned baseline. Four of the 15 fitted level-3 cells have negative $R^2$. Only one of them is large enough to appear in the table: `abdomen / epidermis, dermis / unspecified` ($n = 28$), at $R^2 = -0.120$ despite a below-baseline RMSE of 0.925 — its rows are tightly clustered in $\log_{10} K_p$ (standard deviation 0.87 against 1.27 over the whole in-domain set), so its own mean is a strong competitor and the fit does not beat it. The other three all hold ten rows or fewer, and the worst of them, `abdomen / epidermis / Water` ($n = 10$), reaches $R^2 = -11.0$ at RMSE = 1.69. Both statistics are reported because either alone would mislead.

### 4.4 What we take from this

The claim we are willing to defend from Table 4 is modest and specific. With the descriptors and model form held exactly fixed, error falls as the fit is permitted to condition on more of the experimental context, from a fold-error band of about 18.1 at level 0 to about 14.0 at level 1 to about 9.80 at level 3, and $R^2$ rises from 0.021 0 to 0.373. The fall is not monotonic across

every intermediate step: conditioning on body site alone raises pooled RMSE to 1.16 and lowers pooled $R^2$ to 0.175, both worse than the level-1 refit. Because no molecular information was added, the signal recovered was already present in the corpus as context and was simply not in the model. Because level 3 is computed on 311 of 534 rows, that recovery is suggested, not measured: we cannot rule out that dense protocol cells are easier rows.

The complementary observation is that the largest single obstacle to conditioning is missing context. In each of the three context fields of Table 5, the `unspecified` cell is the worst-performing cell of that field — body site at RMSE = 1.44, skin layer at 2.46, donor solution at 1.34 — and in Table 6 the largest cell, and every cell that fails to beat the level-1 baseline, is one in which the vehicle was not recorded. They are not uniformly the largest cells: the `unspecified` skin-layer cell is the smallest in its field, and across all 15 fitted level-3 cells the single worst RMSE belongs to a ten-row cell with a named vehicle. A database that records substrate and protocol as first-class, queryable structure — in the fiber-bundle model of Section 2, context living on the base space so that a fit can be taken over a fibre rather than over the pooled total space [Davis, 2026a] — is not a modelling nicety here. It is the precondition for the experiment being runnable at all.

That is what motivates the substrate-geometry question this study exists to ask. Fibre geometry — cross-sectional ellipticity, twist density, and the other directly measurable morphological quantities — is a context field that, as far as we have been able to establish, no published permeability model conditions on; we have not run a systematic review and claim no priority. The result above does not show that fibre geometry carries signal. It shows that in this corpus, on this model class, context fields nobody had bothered to condition on carried more signal than the molecular descriptors did, which is the reason to go and measure one.

We state the constraint on that programme explicitly, because it is architectural rather than incidental. Conditioning is on measured geometry only. We do not condition on demographic category, race, or self-reported hair type, and we would not accept a model that did: those are social and self-report variables, and a model stratified on them would be fitting a category assignment while claiming to fit a physical one. We make no empirical claim here about how such labels relate to fibre morphology; establishing that would require data we do not have and have not collected. Ellipticity and twist density are measurable on a fibre; they are the right base-space coordinates precisely because they are.

Finally, the scope of everything in this section. These fits are on pooled public human *skin* permeability data [Stepanov et al., 2020], aggregated across laboratories, protocols, and donors. No hair fibre was measured anywhere in this work. Nothing in Table 4 through Table 6 supports a claim about any product, ingredient, or hair type, and a level-3 fold error of about 9.80 would in any case be far too coarse to support one. What the section establishes is feasibility of a method: hold the model fixed, add context, count the rows you lose, and report the cells that lose.

## 5 Applicability domain and refusal

A prediction is only meaningful where the model has support. This section asks a narrow question with a hard answer: given a representative textured-hair formulation, how many of its ingredients can a descriptor-based permeability model of the Potts–Guy family [Potts and Guy, 1992] answer at all? The answer is that a large minority cannot be answered, and that the unanswerable fraction is concentrated in exactly the ingredient classes that do the structural work in the category.

### 5.1 The domain box

The permeability model at the centre of this paper takes two descriptors, octanol–water partition coefficient $\log P$ and molecular weight $M$, both computed from a single defined molecular structure with RDKit [RDKit, 2026]. We declare its applicability domain as a descriptor box,

$$\mathcal{D} \;=\; \{(M, \log P) \;:\; 18 \le M \le 750 \text{ g/mol}, \quad -3 \le \log P \le 6\}, \tag{7}$$

and apply it identically to the training corpus and to any query, and identically across every analysis reported in this paper. Declaring the domain before scoring, rather than inferring it after the fact from where the model happens to do well, is the practice recommended for QSAR-type models generally [OECD, 2007].

Applied to the ingested huskinDB records [Stepanov et al., 2020], Equation (7) retains 534 of 550 rows; 16 are excluded, 15 of them for $\log P$ and 1 for molecular weight, with no row violating both. The retained corpus realises a narrower envelope than the declared box: $M$ spans 18.02 to 584.66 g/mol and $\log P$ spans $-2.563$ to 5.660. The box is therefore already generous relative to the evidence behind it, which matters below when we measure how far a refused query sits from anything the model has seen.

Two clarifications on scope. First, this is a feasibility study on pooled public human *skin* permeability data; no hair fibre was measured, and nothing in this section supports a claim about any product. Second, the conditioning axes used elsewhere in this paper are recorded experimental-protocol context only — anatomical body site, skin layer and donor solution, as they appear in the source records. We do not condition on demographic category, race, or self-reported hair type: the ingested corpus carries no such field, and a category label is not a measurement.

### 5.2 A representative deck

We assembled a 29-ingredient INCI deck typical of a leave-in conditioning product for textured hair, spanning humectants, acids, scalp actives, preservatives, fatty alcohols and lipids, cationic conditioners, silicones, polymers, protein hydrolysates and a botanical butter. Each entry was classified by a single rule applied in order: if the ingredient has no single defined molecular structure, no descriptors exist and the model is undefined on it; if descriptors exist but fall outside Equation (7), the query is out of domain; otherwise it is in domain. Nothing is imputed. Ingredients without a defined structure are never assigned a representative SMILES to make them scorable.

**Table 7:** Verdicts on a 29-ingredient textured-hair deck under the declared applicability domain of Equation (7).

| Verdict | Ingredients | Share of deck |
|---|---|---|
| In domain: predictable | 17 | 58.6% |
| Out of domain: descriptors outside the box | 4 | 13.8% |
| Undefined: no single defined molecular structure | 8 | 27.6% |
| Refused (out of domain + undefined) | 12 | 41.4% |

The 17 answerable ingredients are the humectants glycerin, propylene glycol, panthenol and urea; lactic and salicylic acid; niacinamide, caffeine and allantoin; the preservative system

phenoxyethanol, benzyl alcohol and ethylhexylglycerin; menthol; cetyl alcohol; caprylic acid; and the two cationic conditioners cetrimonium chloride ($M = 320.00$, $\log P = 3.178$) and behentrimonium chloride ($M = 404.17$, $\log P = 5.519$). Both are admitted by the box, but the support behind them is thin. Behentrimonium chloride sits close to the upper $\log P$ edge and close to the corpus maximum of 5.660: only 1 of the 534 in-domain training records has $\log P$ above 5.5, only 3 exceed $\log P = 5$, and only 35 exceed $M = 400$ g/mol. Both are quaternary ammonium salts entered as two-component structures, and only 7 of the 534 in-domain records are multi-component salts of any kind. Passing a box test is a necessary condition, not a certificate.

### 5.3 What is refused, and why

Table 8 names every refused ingredient with its reason. The two refusal classes are qualitatively different and we keep them separate. The four out-of-domain entries are cases where the model is defined but unsupported: a number can be computed, and it would be an extrapolation. All four violate a $\log P$ bound; none violates the molecular weight bound. The eight undefined entries are cases where the model has no input at all: polymers and mixtures have a distribution of structures, not a structure, so $M$ and $\log P$ are not properties they possess. Descriptor-based prediction here is undefined rather than merely uncertain, and no amount of additional training data changes that.

### 5.4 The silence falls on the category's core

The refusals are not spread evenly across formulation function. Reading Table 8 by role, the refused set is: both conditioning silicones, both cationic polymers, both protein hydrolysates, the botanical butter, the PEG solubiliser, two of the four lipid and fatty-alcohol emollients, the antioxidant, and one humectant, sorbitol. What remains answerable is largely the small polar fraction — the humectants other than sorbitol, the acids, the preservatives, the low-molecular-weight actives — together with the two monomeric quats.

The refused classes are present in a textured-hair product for what they do at the fibre surface: deposition, film formation, charge interaction and lubrication. The answerable classes are, for the most part, the ones a transdermal-permeability model was built for in the first place. So the coverage failure is doubly unfavourable. The model is undefined on the ingredients that carry the product's structural function, and for several of those the quantity it would report — a steady-state skin permeability coefficient — is not the quantity of interest anyway. Even a model that could parse dimethicone would be answering a question about permeation when the formulation question is about deposition. Reported as a single coverage number: the standard model can compute a skin-permeability estimate for 58.6% of this deck and is silent on 41.4% of it, with the silence falling on the part of the category that does the work.

### 5.5 Refusal as the correct output

A descriptor model is a total function on its input space: given any pair $(M, \log P)$ it returns a real number. Nothing in the arithmetic distinguishes an interpolation from an extrapolation. Evaluated at tocopherol's $\log P = 8.840$, the fitted linear model returns a value as readily as it does for glycerin, despite that point lying 3.18 $\log P$ units beyond the largest value in the 534-record training corpus; sorbitol at $\log P = -3.585$ lies 1.022 units below the smallest. Those returned numbers have no measured error bar, because there is no measurement anywhere near them from which an error bar could be estimated. The residual statistics reported elsewhere in this paper are computed inside $\mathcal{D}$ and do not transfer outside it.

**Table 8:** The 12 refused ingredients of the deck. Upper block: descriptors defined but outside the box of Equation (7). Lower block: no single defined molecular structure, so descriptors do not exist.

| INCI name | Role in formula | Reason for refusal |
|---|---|---|
| *Outside the descriptor box* | | |
| Sorbitol | humectant | $\log P = -3.585$; 0.585 below the lower bound |
| Stearyl Alcohol | fatty alcohol / emollient | $\log P = 6.240$; 0.240 above the upper bound |
| Oleic Acid | lipid | $\log P = 6.109$; 0.109 above the upper bound |
| Tocopherol (Vit E) | antioxidant | $\log P = 8.840$; 2.840 above the upper bound |
| *No single defined molecular structure* | | |
| Dimethicone | silicone (polymer) | polymer; descriptors undefined |
| Amodimethicone | conditioning silicone (polymer) | polymer; descriptors undefined |
| Polyquaternium-10 | cationic cellulose polymer | polymer; descriptors undefined |
| Guar Hydroxypropyltrimonium Chloride | cationic guar (polymer) | polymer; descriptors undefined |
| PEG-40 Hydrogenated Castor Oil | PEG ester (polymer) | polymer; descriptors undefined |
| Hydrolyzed Keratin | protein hydrolysate (mixture) | mixture; descriptors undefined |
| Hydrolyzed Wheat Protein | protein hydrolysate (mixture) | mixture; descriptors undefined |
| Shea Butter (Butyrospermum Parkii) | botanical butter (mixture) | mixture; descriptors undefined |

We therefore treat refusal as a first-class output rather than a failure mode, consistent with applicability-domain practice [OECD, 2007]. The check is a property of the model's support, evaluated on the input before any prediction is attempted; it is not a confidence threshold tuned against outcomes, and it does not depend on how the model happens to score. Its two verdicts carry different remedies, which is the practical reason for keeping them apart. Out-of-domain is a data problem: the four refusals in the upper block of Table 8 would become answerable if the corpus were extended to cover high-$\log P$ lipids and, for sorbitol, low-$\log P$ polyols, and the refusal names the missing measurement precisely. Undefined is a representation problem: the eight refusals in the lower block require a different instrument entirely, not more rows of the same kind.

This is also where the fiber-bundle formulation earns its keep operationally. Because a record carries its context explicitly rather than folding it into a fitted value, the domain check can be evaluated on the query's own descriptor coordinates before any number is produced, so a refusal is addressable and auditable — it says which ingredient, which bound, and by how much — instead of being absorbed into a point estimate. Silent extrapolation produces the opposite: a number that is indistinguishable, at the interface, from a supported one. For a category where the best-studied substrates are not the substrate of interest, that indistinguishability is the specific failure worth engineering against.

# 6 Negative results: what we tested and could not establish

We report here a hypothesis that we expected to carry the confidence layer of this study, and that did not survive its own pre-stated test. We report it in full because the failure is informative about what pooled public permeability data can and cannot support.

## 6.1 The hypothesis

GIGI represents a record as a section of a fiber bundle: the base space is the key space, the fiber is the value space. Under this reading, one compound's permeability is not a scalar but a section $\sigma_c : T \to E$ over the experimental context base $T$, where a point of $T$ is a triple (body site, skin layer, donor solution) and $\sigma_c(t)$ is that compound's $\log K_p$ measured in context $t$. The MIRADOR specification assigns each such section a coherence

$$C \;=\; \frac{\tau}{K}, \tag{8}$$

in which $K$ is the Davis curvature of the section as the base flexes and $\tau$ is a fixed tolerance budget [Davis, 2026c]. Davis duality motivates reading $K$ as the obstruction to flat approximation: a section with small curvature is well approximated by a flat model over its base, one with large curvature is not [Davis, 2026a]. The operational hypothesis follows directly and is directional: a compound whose $\log K_p$ curves sharply as the substrate context flexes should be harder for a flat descriptor model to predict, so the rank correlation between $K$ and prediction error should be *positive*.

We estimate $K$ from the section's own measured values by the same quantity GIGI computes per record,

$$K(\sigma_c) \;=\; \frac{\mathrm{Var}\big(\{\log K_{p,i}\}\big)}{\big(\max_i \log K_{p,i} - \min_i \log K_{p,i}\big)^2}, \tag{9}$$

using the population variance. Prediction error is the leave-one-compound-out absolute error of a two-descriptor ordinary least squares model in $\log P$ and molecular weight [Potts and Guy, 1992], with descriptors computed by RDKit [RDKit, 2026], refitted on all other compounds

and asked to predict the held-out compound's mean $\log K_p$. Throughout, $\tau$ is held fixed at 1.0, so $C$ is a strictly decreasing transform of $K$ and every rank statistic we report on $K$ is equivalent up to sign to the same statistic on $C$. We therefore never tested any contribution from $\tau$; that part of Equation (8) is untouched by this study.

### 6.2 The initial result, and the control that undercut it

Restricting to the 534 in-domain records [Stepanov et al., 2020, OECD, 2007] and to compounds with at least three measurements gives 38 sections. Over those sections the Spearman correlation between $K$ and absolute error is $\rho = 0.258$ — the predicted sign. Refusing to answer for the most curved sections appeared to buy accuracy (Table 9): withholding the worst 20 % of sections lowers RMSE from 0.738 to 0.698, an improvement of 5.48 %.

**Table 9:** Refusal by section curvature over the 38 multi-measurement sections. Sections are refused in decreasing order of $K$. The improvement is not monotone in the refusal fraction, which was the first indication that the ordering was not tracking a stable quantity.

| Refused (%) | Sections kept | RMSE | Change (%) |
|---|---|---|---|
| 0 | 38 | 0.738 | – |
| 5 | 37 | 0.728 | 1.34 |
| 10 | 35 | 0.748 | −1.28 |
| 20 | 31 | 0.698 | 5.48 |
| 30 | 27 | 0.711 | 3.62 |

We then ran the control that the construct demands. If $K$ is detecting compounds whose permeability genuinely swings as the substrate flexes, then the raw spread $\max_i \log K_{p,i} - \min_i \log K_{p,i}$ — a cruder but far more direct measure of exactly that instability, and the square root of the denominator of Equation (9) — should correlate with error in the *same* direction, and if anything more strongly. It does not. The Spearman correlation between raw spread and absolute error is $\rho = -0.162$: opposite in sign.

That control mattered because it localises whatever signal $K$ carries into the normalisation rather than into the physics. Two statistics purporting to measure the same instability cannot disagree in sign and both be reading substrate response. The disagreement points at the ratio structure of Equation (9), which is a shape statistic — how variance sits inside range — and is therefore dominated by how many points a section contains and how they are arranged, not by how far apart they are. Across our 38 sections the Spearman correlation between the number of measurements and $K$ is $-0.897$. The six lowest-curvature sections carry between 13 and 16 measurements each (1-Hexanol, $K = 0.048\,3$, $n = 13$; Testosterone, $K = 0.060\,0$, $n = 16$); the six highest carry three or four (Histidine, $K = 0.213$, $n = 3$; Ephedrine, $K = 0.188$, $n = 4$). To first order, $K$ was ranking sample size.

### 6.3 The dose-response test, and the kill

A rank correlation of 0.258 on 38 points is weak evidence at best, so we pre-specified a stronger test than a single $p$-value. The construct is directional, so if it is real the effect should *strengthen* as $K$ is estimated from more repeat measurements per compound. We therefore recomputed $\rho(K, |\text{error}|)$ at increasing minimum-measurement thresholds, with a one-tailed permutation test at each threshold (20 000 label shuffles, fixed seed, and the add-one estimator $\hat{p} = (m+1)/(20\,000+1)$ where $m$ is the number of shuffles with $\rho_{\text{perm}} \geq \rho_{\text{obs}}$, one-tailed because

the theory fixes the sign in advance). This pass runs from the parsed source rows without the applicability-domain filter, which is why the threshold-3 stratum contains 41 sections rather than 38.

The result is Table 10. The correlation does not strengthen. It weakens monotonically and then changes sign, and the permutation $p$ never approaches significance at any threshold.

**Table 10:** Dose-response test of the coherence hypothesis. If curvature carried real information about predictability, $\rho$ should rise as $K$ is estimated from more measurements per compound. It falls and then reverses. The thresholds 5 and 6 select the same 22 sections because no compound has exactly five measurements.

| Min. measurements | Sections | Median measurements | $\rho(K, \lvert\text{error}\rvert)$ | $p$ (one-tailed) |
|---|---|---|---|---|
| 3 | 41 | 6.0 | 0.189 | 0.115 |
| 4 | 32 | 7.0 | 0.158 | 0.188 |
| 5 | 22 | 9.5 | −0.093 2 | 0.664 |
| 6 | 22 | 9.5 | −0.093 2 | 0.664 |

The hypothesis fails its own test. We report no confidence gate built on section curvature, and the refusal curve in Table 9 should be read as noise, not as a usable operating characteristic.

### 6.4 A structural floor: $K$ is uninformative below three measurements

There is a hard reason the thresholds in Table 10 begin at three, and it also explains why the six highest-curvature sections quoted above cluster so tightly, between $K = 0.177$ and $K = 0.213$. With exactly two measurements, Equation (9) is constant. Let $x_1 \le x_2$ with $d = x_2 - x_1 > 0$ and mean $\mu = (x_1 + x_2)/2$. Then $x_1 - \mu = -d/2$ and $x_2 - \mu = +d/2$, so the population variance is $\frac{1}{2}[(d/2)^2 + (d/2)^2] = d^2/4$, while the squared range is $d^2$, and therefore

$$K \;=\; \frac{d^2/4}{d^2} \;=\; \frac{1}{4} \qquad \text{for every two-point section, whatever } d. \tag{10}$$

Two-point sections carry no curvature information at all.

The floor is not confined to $n = 2$. For a section of $n$ points, Equation (9) is confined to

$$\frac{1}{2n} \;\le\; K \;\le\; \frac{1}{4}, \tag{11}$$

the lower bound being attained when two points sit at the extremes and the remaining $n - 2$ sit at the mean. Equation (10) is the degenerate case where the two bounds coincide. At $n = 3$ the attainable interval is $[1/6, 2/9]$ and at $n = 4$ it is $[1/8, 1/4]$, so a three- or four-point section cannot report $K$ below 0.125 whatever its values, while a section with thirteen or more points may go below $1/26$ — and 1-Hexanol, at $K = 0.048\,3$ on $n = 13$, sits just above exactly that floor. A floor falling like $1/(2n)$ is on its own sufficient to produce the $\rho = -0.897$ between sample size and $K$ reported above. Exact ties collapse the interval onto isolated points: Ephedrine's four measurements are three ties at $\log K_p = -5.778$ and one at $-6.326$, and this 3–1 split forces $K = 3/16 = 0.187\,5$ exactly, independently of the gap. Adding two-point compounds to enlarge the sample cannot help, and small-$n$ sections contribute quantisation rather than signal.

A related defect sits in the section definition itself. We required at least three *measurements* but not at least two distinct *contexts*. Four of the 38 sections are measured in a single context and nine in two or fewer, so for those the base does not flex at all and $K$ is measuring replicate scatter within one condition. Any future run of this test must require base variation explicitly.

### 6.5 Diagnosis: context and laboratory are confounded in pooled data

The deeper problem is what "variation across context" means in an aggregated public dataset. The repeat measurements that define a section do not come from one laboratory varying the substrate; they come from different laboratories, published years apart, using different protocols, and pooled after the fact [Stepanov et al., 2020]. Across the 534 in-domain records we count 95 distinct source references. Thirty-one of the 38 sections draw on more than one reference, with a median of four distinct references per section.

We can quantify the confound directly, because some context triples in our data were measured by more than one source. Holding body site, skin layer and donor solution fixed, 18 such multi-source cells occur within the 38 sections, falling in 15 distinct sections. Their internal disagreement is summarised in Table 11: the median gap between the highest and lowest reported $\log K_p$ for the *same compound in the same nominal context* is 0.818 log units, and eight of the 18 cells disagree by at least one full log unit, with a maximum of 3.88.

**Table 11:** Disagreement between source references measuring the same compound in the same nominal context (body site, skin layer, donor solution), within the 38 multi-measurement sections. Gaps are in $\log K_p$ units.

| Quantity | Value |
|---|---|
| Multi-source, fixed-context cells | 18 |
| Sections containing at least one such cell | 15 |
| Median within-cell gap (log units) | 0.818 |
| Maximum within-cell gap (log units) | 3.875 |
| Cells disagreeing by $\geq 1$ log unit | 8 |
| Median (largest within-cell gap) / (section spread) | 0.439 |

A within-cell gap cannot by construction exceed the total spread of the section that contains it, so the honest comparison is a ratio rather than a difference of magnitudes. The median within-cell gap of 0.818 log units sits against a median section spread of 1.891 log units; and across the 15 sections that contain a multi-source cell, the median ratio of the largest fixed-context multi-source gap to that section's entire spread is 0.439. In a typical such section, close to half of what we were calling between-context variation is already reproduced at a single fixed context by different laboratories. The quantity we computed as "curvature of the section as the substrate flexes" is therefore, to a substantial degree, a measurement of how much the contributing laboratories disagreed about that compound. Compounds studied often — the alcohols, the steroids — accumulate many measurements from many groups; the arithmetic of Equation (9) then hands them low $K$, and they are also, independently, the compounds a $\log P$/molecular-weight model fits best. That is a sufficient explanation for the initial $\rho = 0.258$ without any substrate physics.

### 6.6 Untestable, not refuted — and the experiment that would settle it

We are careful about the strength of this conclusion. We have not shown that section coherence fails to predict model reliability. We have shown that pooled public permeability data cannot answer the question: the estimator has a hard information floor below three points, its small-sample behaviour tracks sample size rather than substrate response, and its input variance is dominated by inter-laboratory disagreement rather than by substrate flex. The hypothesis is untestable on this substrate, and we record it as such.

The experiment that would test it is straightforward to state and is a measurement programme, not a reanalysis. One laboratory, one protocol, one analytical method, one vehicle, one compound panel. The substrate is varied deliberately along a measured geometric axis rather than being inherited from whatever each source happened to use, with enough levels spanning the axis — and enough replication at each — that every section clears the structural floor of Equation (11) by a wide margin, rather than sitting at $n = 3$ or $n = 4$ where $K \geq 1/8$ regardless of the data. Genuine replicates inside every (compound, substrate) cell are required so that a variance-components decomposition can separate measurement noise from substrate response before $K$ is computed at all. Only the between-substrate component belongs in Equation (9); our pooled estimate mixed both and could not tell them apart.

Two constraints on that design follow from commitments made elsewhere in this study and we restate them here. First, whatever substrate is used, its axis must be a *measured* physical variable, and never demographic category, race, or self-reported type. We state that as a design rule, not as an empirical finding: nothing in the data analysed here bears on how well any self-reported category tracks a physical substrate variable, and this study is in no position to make that comparison. Second, the whole of the present work, this negative result included, rests on pooled public *in vitro* human *skin* permeability measurements. It is a feasibility study. No hair fiber was measured here; if the programme sketched above is ever run on hair, its substrate axis would be measured fiber geometry — ellipticity, twist density — and that would be a separate experiment on a separate substrate, not an extension of anything measured in this paper. Nothing in this section supports a claim about any product.

### 6.7 A second negative result: the engine's own confidence scorer fails as a refusal gate

Section coherence is not the only thing in this study that failed, and the second failure is recorded here rather than left to the discussion, because it is a failure of our own instrument.

Independently of $K$, we asked whether a confidence score can serve as a refusal gate: rank the 534 in-domain predictions of a $k = 7$ nearest-neighbour model by confidence, refuse the least-confident fraction, and ask whether error on what remains improves. Two scorers were tested, both descriptor-space density measures — a naive inverse mean distance to the retained neighbours, and the engine's own kernel-density confidence over the same three descriptor fields. Both rank error weakly in the desired direction, and the engine's scorer ranks it *better* of the two ($\rho = -0.112$ against $\rho = -0.081\,4$).

On the refusal curve the ordering reverses. Refusing the least-confident 20 % by the naive score improves RMSE from 1.134 to 1.052. The engine's score *degrades* RMSE at every refusal fraction beyond 5 %, reaching 1.156 at 20 % — worse than refusing nothing at all.

A better rank correlation alongside a worse refusal curve is not a contradiction: a rank correlation scores the entire ordering, whereas a refusal gate is judged only on the tail it removes, and the engine's scorer is evidently not ordering that tail usefully here. We draw no architectural conclusion from one dataset, and we neither tuned the scorer for this task nor claim significance for either rank correlation. We report it because a refusal gate that makes predictions worse is exactly the result a system claiming to know its own limits is obliged to publish about itself.

# 7 Discussion

## 7.1 What this study establishes

Three findings survive scrutiny, and all three are about the *database architecture*, not about hair.

First, conditioning on recorded substrate and protocol context reduces prediction error while the descriptor set is held fixed. The same two descriptors — a calculated octanol/water partition coefficient (the Wildman–Crippen $\log P$ estimate) and molecular weight, both computed with RDKit [RDKit, 2026] — give leave-one-out RMSE 1.15 ($R^2 = 0.189$) when fitted on the pooled in-domain corpus with no context, and pooled RMSE 0.991 ($R^2 = 0.373$) when fitted separately within full protocol strata. Against the published Potts–Guy coefficients [Potts and Guy, 1992], which achieve RMSE 1.26 ($R^2 = 0.021\,0$) on this corpus, the fully conditioned fit is 21.2% lower. That total has two parts and we separate them, because only one of them is about context: refitting the same two descriptors on this corpus with no context at all already takes RMSE from 1.26 to 1.15 on the identical 534 rows, a 9.0% reduction, and conditioning then takes 1.15 to 0.991, a further 13.5% relative to that refit — but on the 311 rows full conditioning retains rather than on all 534, so the second step is not a like-for-like comparison (see the third limitation below). Nothing was added to the molecule's representation. The conditioning half of the improvement comes from letting the model know which recorded experimental context — which body site, which skin layer, which donor vehicle — it is predicting for. This is the result that motivates a base-space/fiber layout: context belongs on the base space, where it can index the fit, rather than being averaged away.

Second, a single context axis is not equivalent to the full context. Conditioning on body site alone gives pooled RMSE 1.16 ($R^2 = 0.175$); on skin layer alone, 1.15 ($R^2 = 0.181$); on donor solution alone, 1.10 ($R^2 = 0.274$). Only the joint stratum reaches 0.991. Each of these pooled numbers is computed over the rows its own strata retain — 499, 528 and 441 respectively, against 311 for the joint stratum — so they are not like-for-like either; with that caveat carried, the gain is not attributable to one dominant covariate.

Third, and independently of any model fit, a large fraction of a realistic textured-hair formulation lies outside the domain of the descriptor-based model used here. Of the 29 INCI ingredients we assembled, 8 have no single defined molecular structure at all — two silicones, two protein hydrolysates, two cationic polymers, a PEG ester, and a botanical butter — so molecular weight and partition coefficient are not defined quantities for them, and a further 4 have defined structures that fall outside the applicability box declared for the Potts–Guy model, $18 \leq \mathrm{MW} \leq 750$ and $-3 \leq \log P \leq 6$ [Potts and Guy, 1992], which we adopt as the applicability-domain criterion following OECD guidance [OECD, 2007]. Twelve of 29 ingredients, 41.4%, are therefore unanswerable rather than merely uncertain: eight because no descriptor-based model of any kind can accept an input with no defined structure, four because they fall outside this model's declared domain. A system that returns a number for all 29 is not more capable than one that returns 17 numbers and 12 refusals; it is less honest by exactly 12 answers. Refusal is the correct output for an out-of-domain query, and it has to be a first-class return value rather than an error path.

## 7.2 What this study does not establish

It establishes nothing about hair fibers.

Every measurement analysed in this paper is an *in vitro* human *skin* permeability coefficient drawn from huskinDB [Stepanov et al., 2020]. There is no hair-fiber measurement anywhere in

the corpus, no keratin-fiber permeation assay, no cuticle or cortex endpoint, and no measurement made on a textured fiber of any kind. The strata we conditioned on — body site, skin layer, donor vehicle — are skin-protocol strata. When we write that context conditioning reduces error, the claim is bounded to that: recorded experimental context on pooled human skin data. The extension to hair is a *hypothesis this study was designed to make testable*, not a finding it supports. No sentence in this paper, and no number in it, licenses a claim about how any ingredient behaves on any person's hair, and none licenses a product claim of any kind.

### 7.3 A hypothesis that failed

We tested a second, stronger claim and it did not survive.

The claim was that a compound's reliability can be read off the curvature of its own response as the substrate context flexes. Following the coherence layer of the MIRADOR specification [Davis, 2026c], we treated each compound as a section over the context base and computed, on that section, the dispersion statistic the engine reports as its per-field curvature: $K = \mathrm{Var} \,/\, \mathrm{range}^2$, with Var the population variance. We take this as a low-altitude, finite-sample analogue of the curvature invariants of Davis [2026a] rather than an instance of them; the identification is not established here. We then asked whether $K$ ranks compounds by prediction error. On the 38 in-domain compounds with at least three recorded measurements — sections are admitted on measurement count, not on context count, and 4 of the 38 turn out to have only one distinct protocol context — the rank correlation between $K$ and absolute error is $\rho = 0.258$, in the predicted direction, and better than the naive alternative of raw spread, which correlates at $\rho = -0.162$, i.e. the wrong way.

The discriminator, fixed in advance of the test but not registered anywhere outside this repository, was dose-response: if $K$ is measuring something real, the correlation should strengthen as $K$ is estimated from more measurements per compound. It does the opposite. This test is run over all usable rows rather than the in-domain subset, so its section counts are slightly larger than the 38 above. Requiring at least three measurements gives $\rho = 0.189$ over 41 sections (one-tailed permutation $p = 0.115$); at least four gives $\rho = 0.158$ over 32 sections ($p = 0.188$); at least five or six gives $\rho = -0.093\,2$ over 22 sections ($p = 0.664$) — the last two thresholds select the same 22 sections, because no compound has exactly five measurements. The correlation weakens monotonically as the estimate of $K$ improves and then changes sign. That is the signature of noise, not of a construct, and we report it as a failure of the hypothesis on this dataset.

We note the methodological floor that constrains any retest: for a compound with exactly two measurements, $K = \mathrm{Var} \,/\, \mathrm{range}^2$ computed with the population variance is identically $1/4$, so $K$ carries no information at all below three measurements per section. The failure here is therefore not decisive against the construct in general — a corpus with deeper per-compound context coverage would test it far better — but it is decisive against using it as a confidence signal on data of this shape.

The confidence experiment tells the same story from the other side. It uses a different predictor — a $k = 7$ nearest-neighbour fit over three standardised descriptors (molecular weight, partition coefficient, topological polar surface area), not the two-descriptor linear model above — so its RMSEs are not on the same footing as the numbers reported earlier in this section. Both scorers we tested are descriptor-space density scorers, and both rank error weakly and in the right direction ($\rho = -0.081\,4$ for naive inverse descriptor-space distance, $\rho = -0.112$ for the engine's own kernel-density confidence over the same three fields), but on the refusal curve the naive scorer is the one that pays: refusing its least-confident 20% cuts RMSE from 1.13 to 1.05, a 7.25% improvement, while the engine's scorer degrades RMSE at every refusal fraction

beyond 5%. Descriptor-space density is a weak signal here — weak enough that we claim no significance for either rank correlation — but it is the one that pays on the refusal curve, and the curvature-based alternative of the preceding paragraphs, tested on a different population (38 compound sections rather than 534 records), is not yet a demonstrated improvement on it.

### 7.4 Limitations

1. **Pooled multi-laboratory data.** The 534 in-domain records trace to 95 distinct source references spanning decades. Between-laboratory variance is confounded with every context effect we report; some of the "context" signal is almost certainly laboratory identity. A single-protocol replication is the only way to separate them.

2. **Sparse and inconsistent context metadata.** Context fields are frequently absent. Donor solution is unrecorded for 238 of 534 in-domain records, body site for 81, skin layer for 13. We encode missingness as an explicit `unspecified` level rather than imputing it, which is honest but means several of our largest strata are defined by the *absence* of information.

3. **Thin strata at level 3.** Full protocol conditioning retains only 15 strata with enough rows to fit, covering 311 of 534 records; 223 records — 41.8% of the corpus — fall into cells too thin to fit and are dropped. The pooled level-3 number is therefore computed on a different, easier subset than the level-1 number, and the two are not a like-for-like comparison. The same caveat applies to each level-2 number, which is pooled over the rows its own strata retain. The composition effect is visible in the cells themselves: the largest level-3 stratum ($n = 75$) achieves only $R^2 = 0.022\,9$, while the impressive cells are small ($n = 21$ at $R^2 = 0.837$; $n = 14$ at $R^2 = 0.758$) and carry the optimism that small-$n$ leave-one-out fits always carry.

4. **Negative $R^2$ in some cells.** Conditioning does not help everywhere. Within skin layer, the `dermis` stratum ($n = 33$) gives $R^2 = -0.313$ and the `unspecified` stratum ($n = 13$) gives $R^2 = -0.107$; at level 3, the abdomen/epidermis-dermis/unspecified cell ($n = 28$) gives $R^2 = -0.120$. In those cells the conditioned model is worse than predicting the stratum mean. A conditioning architecture must be able to report this per cell rather than only in aggregate.

5. **No hair-fiber measurements.** Stated again because it is the binding limitation: zero rows in this corpus were measured on hair.

6. **No product-performance endpoint.** Permeability coefficient is not moisture retention, not curl definition, not breakage, not consumer-perceived condition. Even a successful hair-fiber permeability model would be several inferential steps away from anything a formulator ships.

## 8 Proposed panel design

The findings above do not test the OSHUN hypothesis; they establish that the hypothesis is testable and specify the smallest experiment that would test it. We state that experiment here so that it can be criticised before it is run.

## 8.1 Design

*One laboratory, one protocol.* Every measurement made under a single standard operating procedure, single instrument set, single analyst pool. This removes the confound that dominates the present corpus. Protocol is fixed by design, so it cannot be a covariate; only substrate varies.

*Substrate varied on purpose.* Donors recruited to span the fiber-geometry range, not sampled conveniently. The design variable is geometry, and the recruitment target is coverage of the geometry space — specifically, a spread in ellipticity and twist density wide enough that the two are not collinear with each other or with fiber diameter. Collinearity among the geometric covariates is the failure mode that would make the resulting model uninterpretable, and it must be checked at recruitment rather than discovered at analysis.

*Geometry recorded per donor as continuous variables.* Not binned, not classed, not typed. Table 12 lists the covariates to record. Every model covariate in it is a measurement made on the fiber; the two entries that are not — equilibration humidity, which is a chamber condition, and prior chemical treatment, which is a laboratory-recorded exclusion criterion — are marked as such and are not fitted.

*Replicate structure that supports the curvature test.* Each ingredient measured on each donor substrate, so that every ingredient's response is a section over a densely sampled geometry base. The dose-response failure reported in Section 7 was driven by sections whose depth was whatever the literature happened to supply — three to sixteen measurements per compound, median six at the three-measurement threshold — and a panel design fixes this at the source by making the number of contexts per compound a design parameter rather than an accident of the literature.

*Effect size and power.* The target effect is the one observed here: the gap between an unconditioned fit (RMSE 1.15) and a context-conditioned fit (RMSE 0.991) on identical descriptors. Donor count follows from a power calculation against that target; we do not perform that calculation here because its variance inputs must come from single-protocol pilot data, which does not yet exist.

## 8.2 Covariates to record

## 8.3 Ingredient panel

The panel must span the descriptor box rather than cluster in it, and it must deliberately include classes the model is expected to refuse. Concretely: small polar humectants near the low end of the partition-coefficient range, mid-range actives and preservatives, fatty alcohols and lipids at the high end, at least two ingredients outside the box on the partition-coefficient axis in each direction, and at least three of the classes with no single defined molecular structure — a silicone, a protein hydrolysate, a cationic polymer. Of the 29 ingredients assembled here, 17 are inside the box, 4 are outside it, and 8 have no defined structure; that split is close to the right shape for a panel, and it makes the refusal behaviour measurable rather than assumed. Including refused classes is not wasted measurement: it produces the ground truth against which a refusal policy can itself be scored.

## 8.4 Endpoints stated in advance

Table 13 states the endpoints and the decision rule before any data is collected. The primary endpoint is the same class of quantity this study computed on skin and is reported on the

**Table 12:** Covariates to record per donor and per fiber. Every model covariate is a continuous quantity obtained by instrument. No demographic category, no self-reported hair type, and no curl-class label appears in this table, and that is a design requirement rather than an oversight (see the ethics subsection below). The last two rows are not model covariates: equilibration humidity is a controlled chamber condition, and treatment status is a laboratory-recorded exclusion criterion and the only non-continuous entry.

| Covariate | Scale / unit | Rationale |
|---|---|---|
| Cross-sectional ellipticity | dimensionless ratio, minor/major axis | Primary geometric axis; sets the surface-to-volume ratio the ingredient encounters |
| Twist density | twists per cm | Governs local curvature of the fiber surface and therefore contact geometry |
| Major-axis diameter | µm | Scale variable; must be recorded to show geometry effects are not diameter effects |
| Cross-sectional area | $\mu m^2$ | Normalises exposed surface between donors |
| Curl radius, relaxed | mm | Macroscopic consequence of the two microscopic axes; recorded as a check, not a predictor |
| Cuticle scale count per unit length | scales per mm | Barrier-layer density at the point of entry |
| Fiber water content at equilibration | mass fraction | Substrate state at the moment of exposure |
| Equilibration relative humidity | % | Chamber condition, not a fiber measurement; controlled, recorded, and held fixed across donors |
| Prior chemical treatment status | binary, from the laboratory donor record | Exclusion criterion rather than a model covariate; treated fibers are a different substrate |

same scale, so the two sets of numbers can be read side by side; they are not interchangeable, because a hair-fiber coefficient and a skin coefficient are measurements on different substrates.

**Table 13:** Pre-stated endpoints for the proposed panel. Decision rules are fixed before collection so that a negative result is as reportable as a positive one.

| Endpoint | Quantity | Pre-stated decision rule |
|---|---|---|
| Primary | Leave-one-out RMSE of the fixed-descriptor model, unconditioned versus conditioned on measured geometry | Conditioning succeeds only if the conditioned RMSE is lower on the *same* rows, with no stratum dropped |
| Secondary | Per-cell $R^2$, reported for every cell including negative values | Reported in full; no aggregation that hides a negative cell |
| Secondary | Refusal calibration: error on refused versus retained queries | The refusal policy is useful only if refused queries carry higher error than retained ones |
| Secondary | Rank correlation between section curvature and prediction error, with the dose-response check repeated | Declared a failure again if the correlation weakens as sections deepen |
| Negative control | Model conditioned on a randomly permuted geometry label | Must show no improvement; any improvement invalidates the primary result |

### 8.5 Ethics as a design constraint

The model conditions on measured fiber geometry and never on demographic category, race, or self-reported hair type. We state this as a constraint on the architecture rather than as a disclaimer about its use, because it is enforced by what the base space contains: if the categorical field is not recorded, no fit can key on it.

The reason is that the categorical variable is the worse predictor of the quantity the physics depends on. A category label can at best stand as a proxy for fiber geometry, and this study provides no evidence about how good a proxy it is — which is exactly the problem: a model keyed on category would be fitting an unvalidated proxy for the quantity it actually needs, and would then export that unquantified error as a confident per-person prediction. Self-reported hair type has the same defect with an additional one on top: it is a self-report, so its reliability is a property of the reporter and the grading scheme rather than of the fiber. Ellipticity and twist density are the variables the physics depends on, they are measurable to instrument precision, and they are continuous, so they support interpolation between donors instead of forcing every person into the nearest bin.

Correct science and correct ethics select the same architecture here, and it is worth being explicit that this is not a coincidence to be relied upon in general — it is a property of this particular problem, where the ethically fraught variable happens also to be the statistically inferior one. Where the two do come apart, the constraint stands as stated: measured geometry only.

### 8.6 Reproducibility posture

Every number in this paper is produced by a script in the repository from a public, CC-BY dataset [Stepanov et al., 2020]; the confidence curves additionally require a local instance of the engine, whose own confidence scorer is queried over HTTP once per record. The ingest step records provenance, the pull date, and a drop policy under which any row with a missing structure, a missing permeability coefficient, or an unparseable structure is dropped and counted rather than imputed; of 550 raw rows, 550 were usable and 534 fall inside the Potts–Guy applicability box [Potts and Guy, 1992, OECD, 2007], covering 245 distinct compounds. Descriptors are computed from structures with RDKit [RDKit, 2026] rather than transcribed. The conditioning experiment, the coherence experiment, the dose-response test that failed, and the confidence curves are each a single script writing a single JSON file — the confidence script also emits the figure — and the tables in this paper are read from those files. Re-running the scripts against a fresh pull will reproduce the numbers or reveal that the upstream data changed; either outcome is more useful than a number we cannot regenerate.

## 9 Conclusion

This is a feasibility study on pooled public *in vitro* human skin permeability data, and its conclusions are bounded by that.

What is established is narrow and architectural. Recorded experimental context carries information that a two-descriptor structural model does not: with descriptors and model form held exactly fixed, fits taken within protocol strata reach lower leave-one-out error than a global fit of the same form, and no molecular information was added to produce that. The comparison is not made on the same rows, because full conditioning drops the rows that fall in thin or rank-deficient cells, so the effect is demonstrated in direction and only suggested in size. Separately, and independently of any fit, a substantial minority of a realistic textured-hair ingredient deck lies outside what a descriptor-based permeability model can answer. The two reasons for that — descriptors outside the fitted box, and ingredients that possess no descriptors at all — carry different remedies and are reported separately here, because collapsing them into one coverage figure hides the fact that only one of them is fixable with more data. Refusal, on this evidence, is a return value rather than an error path.

What is not established is more. Nothing here concerns hair. Every measurement is human skin from a diffusion-cell preparation, no hair fibre was measured, and no result licenses a claim about any ingredient, product, or person. The coherence construct is not refuted either: pooled multi-laboratory data cannot test it, because the curvature statistic has an information floor at small section depth and its input variance is dominated by inter-laboratory disagreement rather than by substrate response. We record it as untestable on this substrate, not as false. The engine's own confidence scorer was likewise not validated; on this corpus it did not earn a refusal gate. And the largest obstacle we met was not modelling but record-keeping — the strata defined by the *absence* of a recorded vehicle were consistently among the worst-behaved, which is a fact about the literature rather than about skin.

The experiment that would test the substrate hypothesis is stated in Section 8 rather than run, with its decision rules fixed in advance so that a negative outcome is as reportable as a positive one. Stating it before running it is the point.

## References


Bee Rosa Davis. Davis duality: Curvature bounds on flat approximation error, 2026a. Davis Geometric.

Bee Rosa Davis. The double cover principle. Zenodo, 2026b.

Bee Rosa Davis. MIRADOR: Section coherence for candidate reliability, 2026c. Davis Geometric, internal specification.

OECD. Guidance document on the validation of (quantitative) structure–activity relationship [(q)sar] models. Technical Report 69, Organisation for Economic Co-operation and Development, Paris, 2007.

Russell O. Potts and Richard H. Guy. Predicting skin permeability. *Pharmaceutical Research*, 9(5):663–669, 1992. doi: 10.1023/A:1015810312465.

RDKit. RDKit: Open-source cheminformatics. `https://www.rdkit.org`, 2026. Descriptors: `MolWt`, `MolLogP`, `TPSA`. Version as used 2026-08-03.

Dmitri Stepanov, Steven Canipa, and Gerhard Wolber. HuskinDB, a database for skin permeation of xenobiotics. *Scientific Data*, 7:426, 2020. doi: 10.1038/s41597-020-00764-z. CC-BY 4.0.