

Physical Sufficiency and Predictive Identifiability in Amino-Acid Transfer-Energy Representations

Aleksei Novgorodtsev

Independent Researcher, Madrid, Spain

ORCID: 0009-0009-2407-7049 | Correspondence: aleksei.novgorodtsev@gmail.com

Abstract

A molecular representation can preserve every recorded physical response yet fail to determine predictions throughout a validation protocol. We develop an observer-conditioned reduction framework that separates these two preservation requirements and supplies explicit failure certificates.

On the canonical twenty-residue Wimley–White octanol record, a five-coordinate physical representation q_2 is injective and inclusion-minimal within its declared primitive family. Its affine design has rank six and cospark two: every single-residue deletion preserves rank, but deleting cysteine and methionine removes the sulfur direction. Thus finite physical sufficiency coexists with failure of nested leave-one-out predictive identifiability. Exact fiber-loss certificates quantify physical refinement, while sixteen alternative minimal injective encodings rule out uniqueness from injectivity alone. The protocol-specific rank gate reduces 262 low-capacity predictive states to 131; none preserves the complete q_2 state within the tested descriptor grammar.

Fully nested Q^2 values are 0.825, 0.529, 0.586 and 0.213 for octanol transfer, membrane-interface transfer, translocon insertion and the Moon–Fleming scale. Algorithm-matched permutation evidence remains conditional on the retrospectively fixed pipeline. Shared predictive alignment survives a joint-profile-preserving permutation test, without identifying a unique common geometry. An external structure-only FreeSolv demonstration reproduces validation-induced rank loss. Chemotype holdouts delimit extrapolation, and prospective noncanonical validation remains open.

23 Scientific Contribution

24 Physical sufficiency and predictive identifiability require different equivalence relations. The contribution is
25 their integration into an auditable reduction protocol, with a direct finite counterexample to their equivalence.
26 Established projection, thermodynamic-cycle, graph-cycle and sparse-linear-algebra identities supply the
27 certificates; novelty is not assigned to these identities in isolation.

28 **Keywords:** amino acids; transfer free energy; hydrophobicity; thermodynamic contrast; interaction quotient;
29 observer-conditioned state; nested validation; model-space quotient; identifiability

30 1 Introduction

31 Residue-indexed free-energy and hydrophobicity scales are indispensable in membrane biophysics, protein
32 folding, mutation analysis, and molecular design [1,2,3,4,5,6,7,8]. Their common A20 indexing, however, can
33 encourage a stronger interpretation than the experiments warrant: that each residue carries one intrinsic scalar
34 property which is merely measured with different noise. The source experiments considered here do not support
35 that simplification. They differ in carrier, target phase, protonation convention, biological apparatus, and readout
36 construction. A residue label aligns rows; it does not by itself define a complete physical state.

37 Recent work strengthens this premise but also narrows the novelty claim. Redundancy-reduced and ontology-
38 based analyses now organize hundreds of amino-acid scales into nonredundant and semantically classified
39 collections [9,10], underscoring the multiplicity of available residue representations rather than establishing one
40 canonical scale. New experiments distinguish aqueous, membrane-interface, and transmembrane states within a
41 single translocon assay [11]; dewetting-based calculations expose cooperative sequence-context effects in
42 hydration [12]; multiview integration can yield a useful one-dimensional coordinate across distinct water-
43 mediated processes [13]; and computational screening of a million-scale aromatic-centered sequence library
44 shows strong flanking-sequence control of membrane insertion [14]. Context dependence is therefore not itself
45 the unresolved problem. The methodological question is **which molecular distinctions, observer transfers,**
46 **and reductions are licensed once the carrier, state, measurement protocol, uncertainty model, and**
47 **validation domain are declared.**

48 A second problem arises on the modeling side. A low-dimensional regression can be predictive without being
49 physically sufficient, and a syntactically large model space can contain many duplicate predictive states. In
50 small samples, model selection must also be nested inside validation [15,16,17,18,19]. Structural cross-
51 validation has separately highlighted that resampling can expose incompatible fitted structures across folds [20],
52 while contemporary small-data QSAR workflows increasingly emphasize split design, applicability domains,
53 conformal calibration, and reproducible end-to-end pipelines [21]. The present work asks a complementary pre-
54 fit question: whether a candidate predictive representation remains algebraically identifiable on every training
55 domain that the declared validation-and-selection protocol will create.

56 Antibody hydrophobicity benchmarks already show that scale choice, surface scoring and sequence diversity
57 affect agreement with hydrophobic-interaction chromatography [22]. Force-field comparisons of dewetting [23]
58 and transfer energetics in biomolecular condensates [24] further motivate declaring the medium and observable
59 explicitly. These are related physical questions, but neither chromatography retention nor a dewetting score is
60 silently treated here as a WW transfer energy.

61 On the validation side, DataSAIL minimizes similarity leakage between partitions [25], and its subsequent
62 addendum reports benchmark-dependent comparisons with existing splits [26]. Scaffold partitions can still
63 retain substantial molecular similarity [27]. A leakage-aware split and the rank criterion here address different
64 requirements: deployment-relevant separation and identifiability on the training sets actually reached. Passing
65 either requirement does not imply the other.

66 Learned sequence representations such as ProfTrans and ESM [28,29], and mutation-effect benchmarks such as
67 ProteinGym [30], provide a broader context for protein representation learning. Their sequence, structural or
68 fitness targets require their own observation contracts. The present A20 transfer-energy panel supplies neither a
69 matched benchmark of these systems nor evidence of superiority to them. SI S24 maps the closest recent
70 intersections and the contribution retained here.

71 The central claim is that **a representation can be physically sufficient yet predictively inadmissible under a**
72 **specified validation protocol**. Here admissibility means invariant identification of the complete unregularized
73 affine predictive space on the fixed evaluation domain. The present work joins these issues into one ordered

question: **which molecular and predictive distinctions can be removed without changing a declared interaction response, and what type of evidence licenses each removal?** Three technical principles control the answer.

First, transfer response is treated as an **interaction contrast**, not an intrinsic scalar attached to a residue. In a matched phase-by-variant free-energy table, phase-only and variant-only baselines are exact nuisance directions. Second, a new physical coordinate may enter only when a documented fiber counterexample proves a simpler state insufficient. Third, a predictive state is admissible only when it preserves affine rank on every training domain induced by the complete validation and selection protocol.

These rules deliberately allow negative terminals. A representation can be exact on the declared A20 slice yet fail globally; a shared predictor can be useful without identifying a unique geometry; a mechanistic candidate can survive some adversaries while remaining incomplete; and a prospective protocol can be scientifically closed even when its result is still open. The goal is therefore not maximum compression at any cost, but **truth-preserving compression with explicit scope**.

2 Methods and observation contracts

2.1 Molecular possibility space precedes the residue label

Let Ω_{mol} denote a broader admissible molecular-state space containing canonical and noncanonical amino-acid states together with protonation, tautomer/valence, stereochemical, and conformational distinctions. The canonical proteinogenic set A20 is a finite study slice of Ω_{mol} , not the ambient space itself. Molecular distinctions remain distinct unless a target-specific preservation theorem or validated approximation licenses a quotient.

This prevents a common category error: the observed 20-vector is not its own possibility space. The same principle applies to endpoints, descriptors, and observers. We therefore distinguish a broader endpoint-functional space Ω_E , a broader coordinate-rule space Ω_D , and an observer/protocol space Ω_O from the finite E6, D10, and protocol slices analyzed here.

2.2 Endpoint and observer families

The declared E6 slice contains WW octanol, WW interface, Hessa, Moon-Fleming, Eisenberg, and Fauchere-Pliska. The first four are the H1 endpoints fixed for this retrospective audit: WW octanol is the anchor, WW interface is a dependent secondary member of the WW family, and Hessa and Moon-Fleming are additional retrospective campaigns. Eisenberg and Fauchere-Pliska are contextual because their construction and/or units differ from direct transfer free energies. The endpoint identities are summarized in Table 1.

Table 1. Endpoint definitions and scope.

Endpoint	Physical quantity / observer	Unit	Role	State/uncertainty guard
WW Octanol	Whole-residue water-to-octanol transfer free energy	kcal/mol	Anchor	WW state convention; per-residue errors
WW Interface	Whole-residue water-to-membrane-interface transfer free energy	kcal/mol	Dependent secondary	WW family; per-residue errors
Hessa	Sec61/translocon apparent insertion free-energy contribution	kcal/mol	Retrospective	Residue-wise charge state and uncertainty not fully identified
Moon-Fleming	OmpLA-context water-to-bilayer side-chain transfer free energy	kcal/mol	Retrospective	pH 3.8 operational map; row uncertainty audit open
Eisenberg	Normalized consensus hydrophobicity scale	unitless	Context only	Derived consensus, not direct transfer measurement
Fauchere-Pliska	Ac-X-amide partition-derived hydrophobic fragment parameter	unitless	Context only	Carrier/state ledger differs from WW

Observer state is part of the endpoint definition. The current Hessa source contract does not sufficiently identify residue-wise charge state for importing the charge-sensitive coordinate used in the WW anchor. Moon-Fleming is analyzed under the declared pH 3.8 operational state map. Unknown state information is not imputed.

2.3 Analysis provenance and interpretation

All existing endpoint analyses are retrospective. Descriptor formulas can be evaluated from structures without reading the response, but this does not establish that the primitive family, D10 grammar, capacity limit or endpoint set was chosen before first data exposure. These choices are frozen for the final reproducibility audit; no independently timestamped prospective preregistration of their discovery is asserted. Fiber discovery and the decision to refine are response-informed. Closing an endpoint during coordinate evaluation does not erase earlier exposure.

The permutation campaigns repeat the declared computational selection procedure, conditional on its fixed inputs and design choices. They do not repeat or correct the preceding human search over possible feature universes, capacity limits, endpoints or coefficient templates. Prospective contracts concern future untouched measurements only. Supporting Information (Additional file 1), Section S17, retains the full claim ledger; the main text separates exact statements, finite-ledger certificates, retrospective predictive evidence and untested physical extensions.

Figure 1 summarizes the two reduction branches.

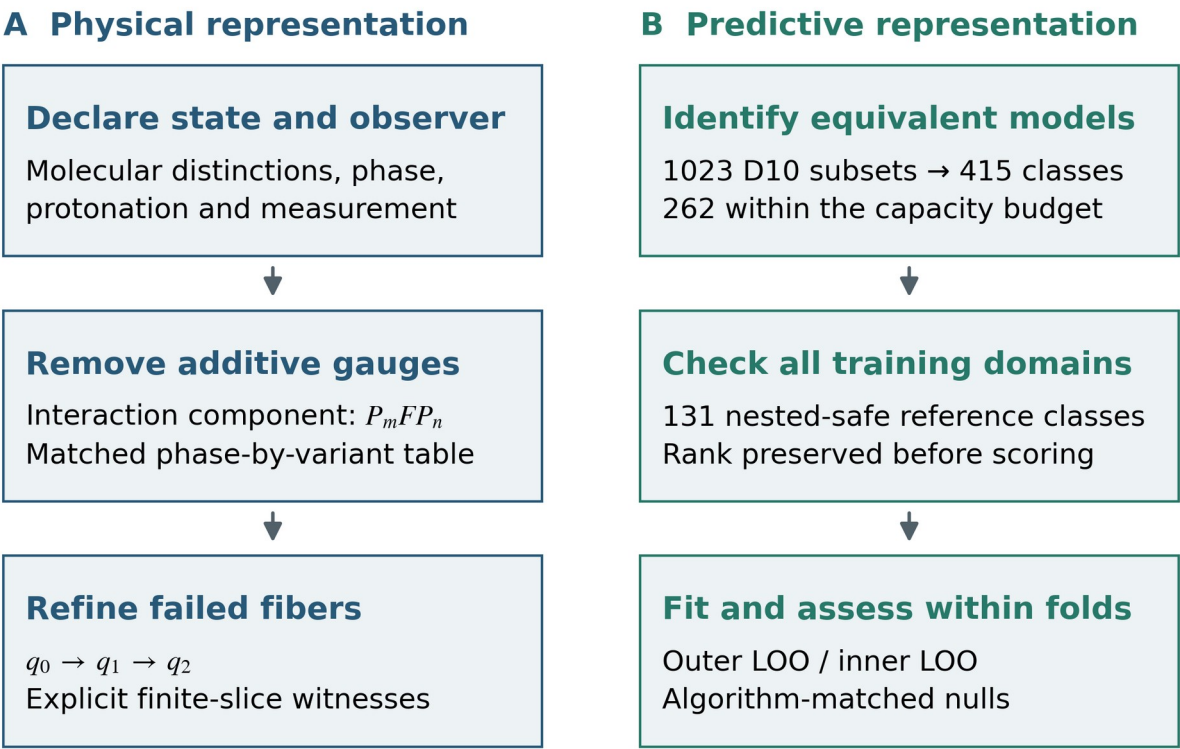


Figure 1. Physical and predictive reduction. The physical branch declares the state and observer, removes additive free-energy gauges and refines failed fibers. The predictive branch quotients affine-equivalent models and checks every training domain before nested fitting. The two branches preserve different objects; response values do not define upstream structural admissibility.

126 3 Theory and representation

127 3.1 Thermodynamic interaction quotient

128 3.1.1 Free-energy contrasts quotient absolute gauges

129 Consider a declared equilibrium ensemble with partition function Z and free energy $F = -\beta^{-1} \log Z$. Any
 130 experimental observable built from a free-energy difference removes an additive absolute-free-energy gauge.
 131 This is an exact observer-level reduction; it does not identify the microscopic Hamiltonian because many
 132 microscopic models can yield the same partition-function ratio.

133 For a vector $F \in \mathbb{R}^m$ of macrostate free energies, the contrast map has kernel $\text{span}\{1\}$ and rank $m - 1$. Thus
 134 the exact macrostate contrast state is $\mathbb{R}^m / \text{span}\{1\} \cong 1^\perp$.

135 3.1.2 Phase-by-variant interaction quotient

136 Let $F \in \mathbb{R}^{m \times n}$ be a matched table indexed by environment/phase and molecular variant. Define the additive
 137 nuisance subspace

$$138 \quad A = \{a 1_n^T + 1_m b^T\}. \quad (1)$$

139 The map $(a, b) \mapsto a 1_n^T + 1_m b^T$ has a one-dimensional kernel, hence $\dim A = m + n - 1$. With

$$140 \quad P_m = I_m - \frac{1}{m} 1_m 1_m^T, P_n = I_n - \frac{1}{n} 1_n 1_n^T, \quad (2)$$

141 the exact mixed interaction component is

$$142 \quad F_{int} = P_m F P_n. \quad (3)$$

143 Its image is $1_m^\perp \otimes 1_n^\perp$ with dimension $(m - 1)(n - 1)$, while its kernel is exactly A . Therefore

$$144 \quad \mathbb{R}^{m \times n} / A \cong 1_m^\perp \otimes 1_n^\perp. \quad (4)$$

145 For two phases ϕ, ψ and two molecular variants x, y , the elementary witness

$$146 \quad \Gamma_{\phi\psi;xy} = F_{\phi x} - F_{\phi y} - F_{\psi x} + F_{\psi y} \quad (5)$$

annihilates all phase-only and variant-only offsets. Conversely, all such double differences vanish if and only if the table is additive. In the equilibrium partition-function representation,

$$\Gamma_{\phi\psi;x y} = -\beta^{-1} \log \frac{Z_{\phi x} Z_{\psi y}}{Z_{\phi y} Z_{\psi x}}. \quad (6)$$

The elementary Γ contrast is algebraically analogous to the coupling contrast used in thermodynamic double-mutant cycles, where a four-state free-energy cycle isolates nonadditivity between two perturbation axes [31]. Here the axes are environment and molecular variant rather than two mutations. The doubly centered matrix $P_m F P_n$ packages the complete phase-by-variant interaction component while quotienting all additive row-only and column-only gauges. This lineage separates familiar alternating free-energy algebra from the present contribution: a typed observer-level quotient embedded inside a larger reduction-and-validation protocol. This is the exact observer-level physical target used in the reduction protocol. It says what a matched transfer family identifies at the observer level. It does **not** specify a unique microscopic channel basis.

3.2 Physical representation and fiber refinement

3.2.1 Response-blind reference and fiber rule

A projection q is insufficient for a declared response if two states share the same coordinate but have different responses: $q(\omega) = q(\omega')$ while $Y(\omega) \neq Y(\omega')$. Such a failed fiber **forces refinement of the induced state partition, not a unique refining coordinate**. Within the declared primitive family used here, one new response-blind coordinate is admitted at a time only after a failed fiber is documented; the resulting coordinate is then fixed for the replay. This audit ordering is not evidence of independent preregistration.

The response-blind first-order physical reference is

$$q_0 = (n_{CS}, n_{ON}, b_{Born}), b_{Born} = \frac{Z_{formal}^2}{\sqrt{A_{polar}}}. \quad (7)$$

Here A_{polar} is the fragment-based topological polar surface area (TPSA, Å²) of the frozen pH-7 side-chain analogue, using the fragment convention of Ertl et al. [32]. Thus b_{Born} has units Å⁻¹. For neutral states, both

169 z_{formal} and b_{Born} are zero. In the machine-readable physical-state ledger this same descriptor is stored under the
 170 implementation label Born_TPSA; that field is b_{Born} , not A_{polar} itself.

171 On the declared A20 state ledger, q_0 merges the exact fibers $\{I, L, M\}$ and $\{P, V\}$. The first declared witness
 172 L/I is separated by the binary beta-branching indicator β_{br} , producing $q_1 = (q_0, \beta_{br})$. The only remaining
 173 nontrivial fiber is then L/M . The WWO record for that fiber gives $\Delta G(L) = -1.25$ and $\Delta G(M) = -0.67$
 174 kcal/mol, a nonzero defect of 0.58 kcal/mol. The atom-type distinction selected from the declared primitive
 175 family is sulfur presence s . The final finite state is

$$176 \quad q_2 = (n_{CS}, n_{ON}, b_{Born}, \beta_{br}, s). \quad (8)$$

177 3.2.2 Exact finite closure and minimality

178 On the declared A20 WWO slice, q_2 is injective. Endpoint constancy on its fibers is therefore exact by
 179 injectivity. This is a finite-slice certificate, not a claim that the microscopic response has five dimensions.

180 An exhaustive audit of all $2^5 = 32$ subsets of the five primitive coordinates shows that the full five-coordinate
 181 state is the **only** exact WWO point-estimate-sufficient subset inside this declared primitive family. Deleting any
 182 one coordinate creates an explicit failed fiber: F/G for n_{CS} (2.86 kcal/mol), H/L for n_{ON} (1.36), E/Q for b_{Born}
 183 (2.86), P/V for β_{br} (0.60), and L/M for sulfur presence (0.58). The rationalized 20×5 coordinate matrix has
 184 exact column and affine rank five.

185 A response-blind negative control makes the nonuniqueness explicit. Exhaustive enumeration of the
 186 independent D10 descriptor ledger finds that five coordinates are also necessary for an injective D10 encoding
 187 of A20, but **16 distinct five-coordinate D10 subsets** achieve injectivity. Finite injectivity therefore does not
 188 identify a unique molecular geometry. The status of q_2 comes from its declared physical primitive family and
 189 witness-driven refinement path, not from injectivity alone (SI S7.1).

190 Figure 2 connects the physical refinement path to its predictive rank boundary.

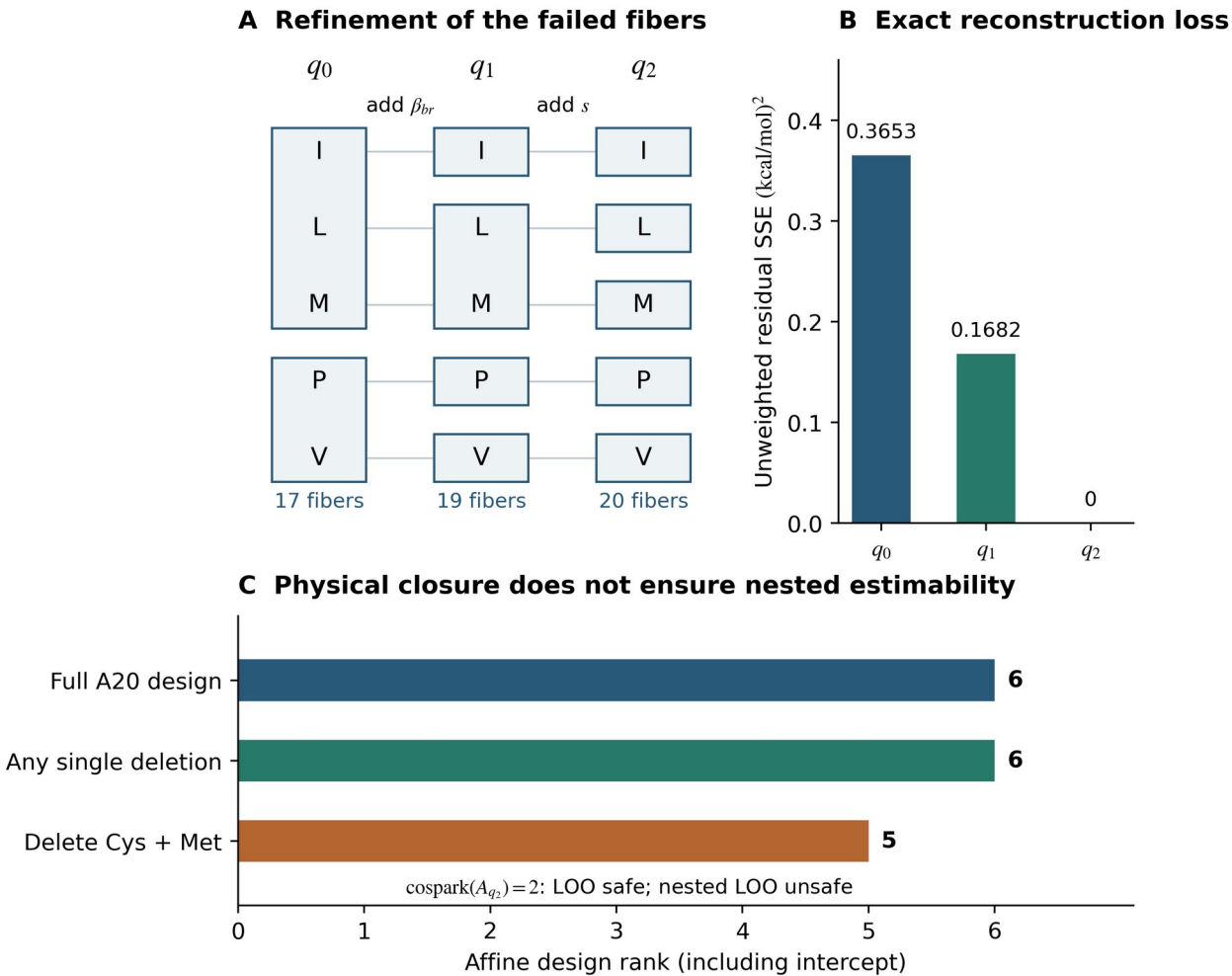


Figure 2. Witness-driven physical refinement. A, Branching separates I from L/M and resolves P/V; sulfur presence separates L/M. The other 15 residues remain singletons. B, The unweighted residual sum of squares decreases from 5479/15000 through 841/5000 to zero. C, The q_2 affine design retains rank after every single deletion, but loses the sulfur direction when Cys and Met are removed together, the only failing pair among 190 possible pair deletions. Refinements belong to the declared primitive family and do not identify a unique physical geometry.

3.2.3 Rank firewall

None of these finite facts identifies intrinsic response rank. A scalar WWO output can be reproduced from latent representations of sample rank 1 through 20 by choosing a fixed scalar observer that reads one coordinate. Hence one scalar endpoint does not identify latent physical rank. Four quantities must remain separate: observer codomain dimension 1, q_2 structural rank 5, exact WWO record quotient size 20, and intrinsic response rank **NONIDENTIFIABLE**.

3.2.4 Quantifying the cost of a failed fiber

Fiber failure admits a quantitative certificate. For a state-by-response table R , let Π_q average within the fibers of q and define $\Delta_q = \|R - \Pi_q R\|_F$. This is the minimum Frobenius reconstruction error over all functions of q on the finite record, including nonlinear functions. Its magnitude depends on response units, weights and the chosen norm; it is not a universal information-theoretic quantity. If q' refines q , orthogonality gives

$$\Delta_q^2 = \Delta_{q'}^2 + \|(\Pi_{q'} - \Pi_q)R\|_F^2. \quad (9)$$

For WW octanol, the squared defects of q_0, q_1, q_2 are exactly 5479/15000, 841/5000 and zero $(kcal/mol)^2$. Relative to the norm of the centered endpoint, the defects are 8.656%, 5.874% and zero. These are norm ratios, not explained-variance percentages. The corresponding best possible uniform errors are 0.30, 0.29 and zero kcal/mol (SI S19). Thus refinement has a measurable finite-record benefit, while injective closure still supplies no guarantee of prediction beyond that record. Chemical definitions are response-blind; the discovery decision to refine a failed fiber is response-informed.

3.3 Transfer across observers

The same residue-side-chain structural coordinate can remain meaningful when the observer changes, while the physical experiment does not transfer. Conversely, two assays can share a log-ratio algebra without sharing a molecular state. We therefore type state transfer and observer/protocol transfer separately (Table 2).

Table 2. Observer and state transfer boundaries.

Endpoint	q2 state transfer	Observer/protocol transfer	Main boundary
WW Octanol	Transfer	Transfer	Reference identity
WW Interface	Transfer	Partial transfer	Same host-guest/state family; target phase changes
Hessa	Not identified	No transfer	Charge-state map not source-identified; Sec61 apparent-insertion observer
Moon-Fleming	No transfer	No transfer	pH 3.8 charge map and OmpLA folding observer differ
Fauchere-Pliska	Not identified	Partial transfer	Same water/octanol contrast layer, different carrier/state ledger
Eisenberg	No transfer	No transfer	Derived consensus scale, no single direct observer

WW interface is the strongest physical relation to the anchor: the host-guest carrier and selected state convention transfer, while the target phase changes. Hessa lacks a source-identified residue-wise charge-state map for

222 importing the full q2 state and uses a Sec61/glycosylation-derived apparent insertion observer. Moon-Fleming
 223 uses an OmpLA folding context with a pH 3.8 state map incompatible with the WW pH-7 charge-sensitive
 224 coordinate. Fauchere-Pliska shares the water/octanol phase pair and relative partition algebra but not the same
 225 carrier or closed protonation ledger. Eisenberg is a derived consensus scale rather than one direct physical
 226 observer.

227 The consequence is strict: **shared residue labels do not establish a common physical observer**. Cross-
 228 endpoint predictive sharing must therefore be interpreted as shared predictive organization unless an upstream
 229 transfer theorem is already available.

230 3.4 Predictive equivalence and validation safety

231 3.4.1 Semantic rather than syntactic capacity

232 The declared D10 grammar contains ten response-blind descriptor rules. For affine OLS, two syntax subsets are
 233 semantically equivalent if they generate the same affine predictive column space on the declared finite domain.
 234 The 1023 nonempty D10 syntax subsets collapse to 415 semantic classes. Two engineered exact redundancy
 235 controls make this a positive control: when both controls are removed, 255 syntax subsets map to 255 semantic
 236 classes. No hidden chemical redundancy is inferred from the engineered collapse.

237 Semantic capacity is $d_{eff} = \text{rank} \left(\begin{bmatrix} 1 & X \end{bmatrix} \right) - 1$. The low-capacity budget is $d_{eff} \leq 4$.

238 3.4.2 Validation-depth theorem

239 Let $A = \begin{bmatrix} 1 & X \end{bmatrix}$ be a full-column-rank affine design and let $T(V)$ be the training-row domains actually generated
 240 by a complete validation/selection protocol V . A predictive state is protocol-safe if and only if

$$241 \quad \text{rank}(A_T) = \text{rank}(A) \text{ for every } T \in T(V). \quad (10)$$

242 Equivalently, for a deletion set D ,

$$243 \quad \text{rank}(A_{-D}) < \text{rank}(A) \quad (11)$$

244 if and only if there exists $c \neq 0$ with $\text{supp}(Ac) \subseteq D$. Thus rank failure occurs precisely when a nonzero
 245 predictive direction is supported entirely within a deletion set reached by the protocol.

246 For outer-LOO/inner-LOO, every unordered two-residue deletion is reached. The reference D10 grammar
 247 contains 262 low-capacity semantic states, of which 131 remain validation-depth safe. The observer-safe Hessa
 248 grammar retains 47 of 93 low-capacity states, and the Moon pH-3.8 grammar retains 89 of 161.

249 The deletion theorem has a standard sparse-linear-algebra invariant. For a full-column-rank affine design A ,
 250 define

$$251 \quad \text{cospark}(A) = \min_{c \neq 0} \|Ac\|_0, \quad (12)$$

252 the minimum support size of a nonzero vector in the column space [33]. The support characterization above
 253 gives

$$254 \quad \text{rank}(A_{-D}) = \text{rank}(A) \forall |D| \leq d \Leftrightarrow \text{cospark}(A) > d. \quad (13)$$

255 Hence the present nested-LOO criterion is equivalently $\text{cospark}(A) > 2$, and exact enumeration certifies this
 256 inequality for each of the 131 surviving designs out of 262. This scalar invariant does **not** replace the general
 257 protocol-specific rule: scaffold, group, cluster, or chemotype splits generate structured deletion families $T(V)$
 258 that need not contain every subset up to a given cardinality. Structural cross-validation addresses a related
 259 problem of fold-dependent fitted structure [20]; here the representation itself is screened for exact affine-rank
 260 preservation before target scoring.

261 For a particular held-out row a , the weaker necessary and sufficient condition is $a \in \text{rowspan}(A_T)$. Full rank
 262 preservation ensures that every direction of the complete finite evaluation design is supported. Ridge or another
 263 regularizer may legitimately define a prediction after rank loss, but it adds a predictor convention; it does not
 264 restore identification of the original unregularized quotient. A penalty must transform with a change of basis if
 265 the same regularized predictor is intended (SI S20).

266 This gate is applied **before target scoring** within the fixed retrospective grammar. Cross-validation therefore
 267 constrains the admissible predictive-state space, not merely the error estimate.

268 3.4.3 An explicit unsupported-prediction witness

269 The rank condition prevents a concrete ambiguity. In the charge-family control design $[1, n_{CS}, n_{ON}, f_{DEK}, f_R]$,
 270 omitting the single Arg row reduces rank from five to four. Ordinary pseudoinverse LOO reproduces
 271 $Q^2 = 0.92417$, but replacing f_R by $1 + f_R$ changes it to 0.93813 despite preserving the full affine column space.
 272 The held-out Arg prediction changes from 0.64963 to 0.99401 kcal/mol. This is a dependence on an
 273 unidentifiable training-null direction, resolved differently by two coordinate-dependent minimum-norm
 274 conventions (SI S20). It explains why an executable LOO score does not alone certify an affine predictive
 275 quotient. The example is separate from the frozen nested-safe H1 campaign.

276 3.5 Direct separation of physical sufficiency and predictive identifiability

277 The physical representation itself gives a direct validation-depth certificate. For $A_{q_2} = [1, q_2]$, exact rational
 278 elimination on the recorded ledger gives rank six. All 20 single-row deletions preserve it; among all 190 pairs,
 279 only deletion of Cys and Met reduces rank to five. The sulfur column is supported exactly on these two rows.
 280 Consequently,

$$281 \quad \text{cospark}(A_{q_2}) = 2. \quad (14)$$

282 Ordinary LOO is therefore rank-safe for this design, while outer-LOO/inner-LOO is not. This failure is
 283 independent of the predictive capacity limit: q_2 has five affine coordinates and would still fail nested rank safety
 284 if the four-coordinate budget were relaxed. Injectivity allows arbitrary finite-record reconstruction; it does not
 285 imply exact affine fitting or estimability on smaller training sets.

286 Both full D10 and q_2 separate all twenty declared A20 labels, but they are not affine-equivalent physical
 287 representations. The full D10 affine dimension is eight, q_2 affine dimension is five, and b_{Born} lies outside the
 288 full D10 affine span on the fixed ledger. Sulfur is represented in D10 through an engineered sulfur coordinate,
 289 yet its A20 support is only C and M; under the two-deletion validation protocol this sparse direction cannot
 290 survive as an independent nested-safe coordinate.

Within the declared D10 grammar and capacity budget, no single nested-safe reference class is injective on all A20 labels: the best safe classes separate at most 19 rows. No nested-safe reference class preserves b_{Born} or sulfur presence affinely, while five safe classes preserve the three-coordinate structural core $(n_{CS}, n_{ON}, \beta_{br})$. Therefore the physical state and the validation-safe predictive state are **different quotients with different losses**. Exact physical sufficiency on the fixed record is not the same claim as validation-safe predictive sufficiency.

This impossibility is restricted to the tested affine family. An arbitrary scalar code could distinguish twenty labels, so no universal impossibility of finite encoding is claimed. Physical primitivity, finite injectivity and fold-wise estimability answer different questions.

4 Predictive results

4.1 H1: endpoint-specific compression

The H1 protocol uses outer leave-one-residue-out validation and literal inner LOO on each 19-residue training set. Candidate classes are limited to semantic states that already passed the validation-depth gate. Selection uses inner MAE, then RMSE, then lower d_{eff} and deterministic class id. Every response permutation reruns the identical adaptive selection algorithm.

Throughout the predictive audit,

$$Q^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \quad (15)$$

where $\hat{y}_i$ is the prediction for residue i from the corresponding outer held-out fold and $\bar{y}$ is the mean observed response over the evaluation panel.

Table 3 reports endpoint performance and the two null tests.

Table 3. Fully nested endpoint results and permutation tests.

Endpoint	Candidate family	Safe n	Q^2	MAE	RMSE	Holm p	Stress p
WW Octanol	D10	131	0.825	0.512	0.654	8e-05	2e-05
WW Interface	D10	131	0.529	0.442	0.573	0.00116	0.00266
Hessa	D8 observer-safe	47	0.586	0.600	0.821	0.00066	0.00072
Moon-Fleming	D10 pH 3.8	89	0.213	1.463	2.001	0.0229	0.113

MAE and RMSE are in kcal/mol. Holm p is adjusted across the four algorithm-matched endpoint tests; stress p is the unadjusted auxiliary best-fixed-class null result.

All four H1 endpoint campaigns fixed for the retrospective audit clear the algorithm-matched 50,000-permutation null after Holm correction. Moon-Fleming, however, does not receive the same terminal strength as the other three: its auxiliary best-fixed-class stress-null p-value is 0.113, its outer folds select seven distinct classes, and the post hoc modal-class full-LOO Q^2 exceeds the fully nested value by 0.397. Row-level uncertainty provenance also remains open. Its terminal status is therefore conditional.

Figure 3 summarizes the nested predictive scores and their statistical qualifications.

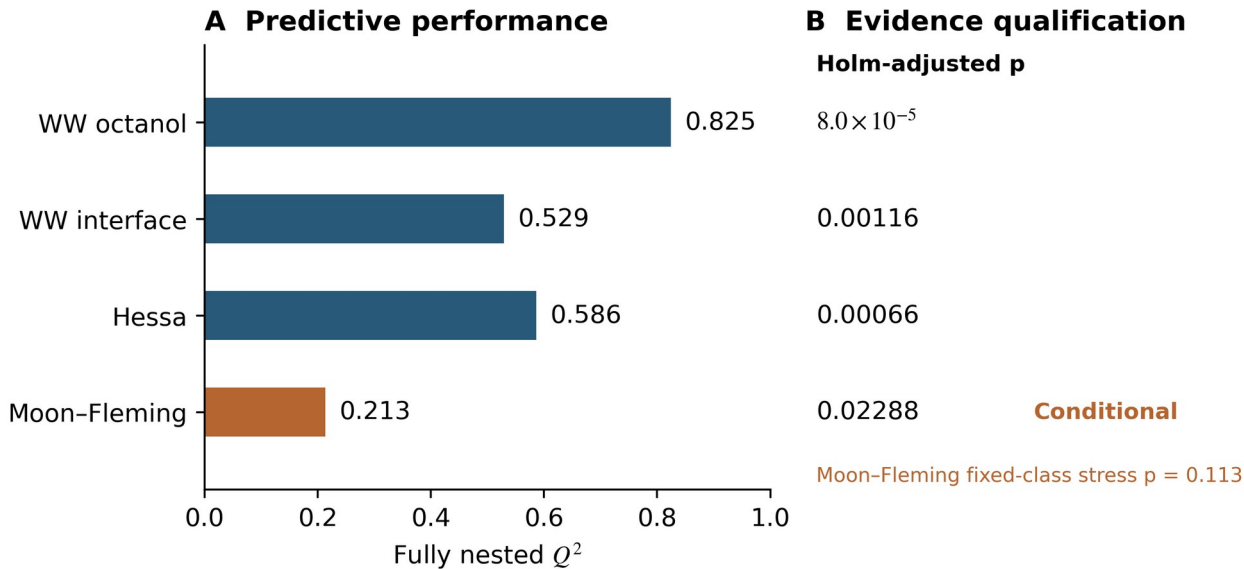


Figure 3. Fully nested endpoint prediction. A, Bars show outer-LOO/inner-LOO Q^2 . B, Holm-adjusted p-values come from the algorithm-matched 50,000-permutation null. Moon-Fleming remains conditional because the auxiliary fixed-class stress test gives p = 0.113 and model selection is dispersed.

324 4.2 Shared prediction is not shared geometry

325 Using the intersection observer-safe grammar, 47 classes remain available for shared-model selection with
 326 endpoint-specific coefficients. A four-endpoint dependency-weighted campaign gives aggregate $Q^2 = 0.5491$; a
 327 dependency-reduced three-endpoint campaign gives $Q^2 = 0.5756$. In both campaigns, 0 of 10,000 independently
 328 sampled within-endpoint residue-label permutations reach the observed aggregate statistic (plus-one $p \approx 10^{-4}$).
 329 Independent endpoint permutations destroy empirical cross-endpoint dependence. We therefore additionally
 330 apply one common residue permutation to every endpoint, preserving whole response profiles and their
 331 covariance, and repeat the identical nested shared selector. Both campaigns again give zero exceedances in
 332 10,000 draws: plus-one $p = 1 / 10001$ and Holm-adjusted $p = 2 / 10001$ across these two tests (SI S10).
 333 Synchronized permutations are established methodology [34]. The tested null is exchangeability of whole
 334 residue profiles relative to descriptor rows; this is not a null conditioned on chemical strata.

335 The evidence supports alignment of joint endpoint profiles with a useful shared predictive class. It does not
 336 establish superiority to separate predictors or identify a unique common generator. The three-endpoint campaign
 337 selects $(n_{CS}, n_{ON}, \beta_{br})$ in all 20 outer folds, but heavily overlapping folds are not independent replications. A
 338 response-conditioned local sensitivity simulation remains illustrative only (SI S10).

339 4.3 Comparator and applicability boundaries

340 The predictive grammar is not universally superior. Matched-grammar Ridge, PCR, and PLS baselines were
 341 fitted on response-blind linearly independent bases selected in the frozen descriptor order derived from each
 342 endpoint's frozen grammar, with scaling, tuning, and fitting confined to the same nested training folds. For WW
 343 octanol, H1 gives $Q^2 = 0.825$ versus 0.738/0.583/0.721 for Ridge/PCR/PLS; for WW interface, 0.529 versus
 344 0.612/0.575/0.708; for Hessa, 0.586 versus 0.374/0.416/0.367; and for Moon-Fleming, 0.213 versus
 345 0.432/0.185/0.295. H1 is strongest for WW octanol and Hessa, whereas PLS is stronger for WW interface and
 346 Ridge is stronger for Moon-Fleming. The contribution is therefore validation-safe state-space construction and
 347 auditable compression, not endpoint-universal predictive dominance.

348 This division of labor is consistent with current small-data QSAR practice. ProQSAR, for example, emphasizes
349 reproducible modular pipelines, scaffold/cluster-aware splitting, conformal calibration, and applicability-domain
350 assessment [21]. Those safeguards operate downstream of molecular representation; the validation-depth gate
351 here is complementary because it asks whether the representation remains identifiable on every training domain
352 before response-based selection begins.

353 Broader Generic11 Ridge gives an additional external comparator boundary. For WW octanol specifically,
354 adding the formal-charge magnitude $|z_{formal}|$ to Generic11 Ridge raises Q^2 from 0.249 to 0.817, close to the
355 compression audit's 0.825. Together these controls explicitly reject a universal predictive-superiority claim.

356 Broad unseen-chemotype extrapolation is also not certified. Leave-one-chemotype-out gives $Q^2 = -1.977$ for
357 WW octanol and -1.356 for WW interface, while Hessa and Moon-Fleming remain positive at 0.277 and 0.289.
358 For WW octanol, acidic and basic holdouts contribute 95.29% of stress SSE. A charge-free observer-safe rerun
359 gives $Q^2 = 0.232$ for WW octanol, indicating that much of the failure is localized to sparsely supported charged-
360 family structure without identifying its cause or establishing broad extrapolative validity.

361 4.4 External structure-only demonstration

362 A response-free audit of FreeSolv molecular structures [35] tests whether the failure is specific to amino acids.
363 Using an intercept and counts of C, N, O, S, F, Cl, Br and I, the complete 642-molecule design and a fixed 40-
364 molecule subset both have rank nine and preserve rank under every ordinary LOO deletion. Holding out each
365 element-containing family instead lowers rank: to eight in the full panel and to seven or eight in the small panel.
366 Every held-out row is then outside the training row space. Invertible basis changes alter the resulting minimum-
367 norm prediction operators.

368 The intercept-plus-carbon-count control retains rank two and invariant prediction operators on every same
369 family split. The 40 molecules were selected by numeric database identifier, without response values or
370 replacement based on outcome. These deliberately diagnostic family splits expose loss of support; they are not
371 an estimate of how frequently the problem occurs in routine splits. SI S25 supplies all structures, exact ranks and
372 operator checks. No hydration-energy prediction or improvement claim is made.

373 5 Discussion and scope

374 5.1 Predictive stability does not identify a structural explanation

375 Under 10,000 bounded WW uncertainty perturbations, octanol Q^2 has 2.5% and 97.5% quantiles 0.796 and
376 0.851, but the original modal class is retained in only 34.73% of draws (SI S15). Aggregate predictive
377 performance can therefore be comparatively stable while selected model identity changes. These perturbation
378 quantiles are sensitivity summaries, not confidence intervals, and do not establish stability of each individual
379 held-out prediction.

380 Hessa selects the same structural class in 19 of 20 folds. That internal consistency is informative, but its
381 incomplete residue-wise uncertainty ledger prevents a corresponding uncertainty-stability claim. Neither
382 repeated fold selection nor a good score identifies physical coordinates uniquely.

383 5.2 Physical models and mechanistic tests

384 The Born-like coordinate is a physically motivated amplitude proxy with finite-record consequences and a
385 frozen prospective competitor test. It is not a first-principles mechanism. Conserved-support geometry addresses
386 a different candidate object; its metric, carrier and solvent-response bridges remain unvalidated. SI S13 and S16
387 retain the mathematical construction and crossed experimental design as a future discrimination program.

388 The historical coefficient coincidence is retained solely as a cautionary case in SI S12 and S22. Its conditional
389 fixed-basis permutation result does not establish intrinsic dimension, does not correct response-exposed design
390 choices, and does not transfer to the chemically corrected Cys coding.

391 5.3 Chemistry and environment controls

392 The survivor set was challenged without tuning across charge, carboxylate/guanidinium, protonation, tautomer,
393 conformation, environment, chemotype, and observer-deletion axes.

394 Protonation cannot be globally quotiented out: the retained WW-interface neutral-His value is 0.17 kcal/mol
395 while the historical charged-His row is 0.96, a 0.79 kcal/mol state sensitivity control. Tautomer sufficiency is

not identified because no tautomer-resolved endpoint pair is available. Conformers can be marginalized exactly inside the equilibrium partition function for the free-energy target, but a static residue-level coordinate is not thereby a conformer-complete microscopic mechanism.

The closest environment adversary is WW octanol versus WW interface because the host-guest family is shared. If the environment change were only a residue-independent offset, $d_x = W W I(x) - W W O(x)$ would be constant. Instead its range is 3.22 kcal/mol, population SD is 0.858 kcal/mol, and the largest centered deviation is 2.1025 kcal/mol at Asp, even though the raw endpoint correlation is 0.9194. Therefore a residue-independent environment-offset model is a **NO-GO**; the mixed molecule-by-environment interaction remains necessary.

Figure 4 shows the residue-wise environment contrast.

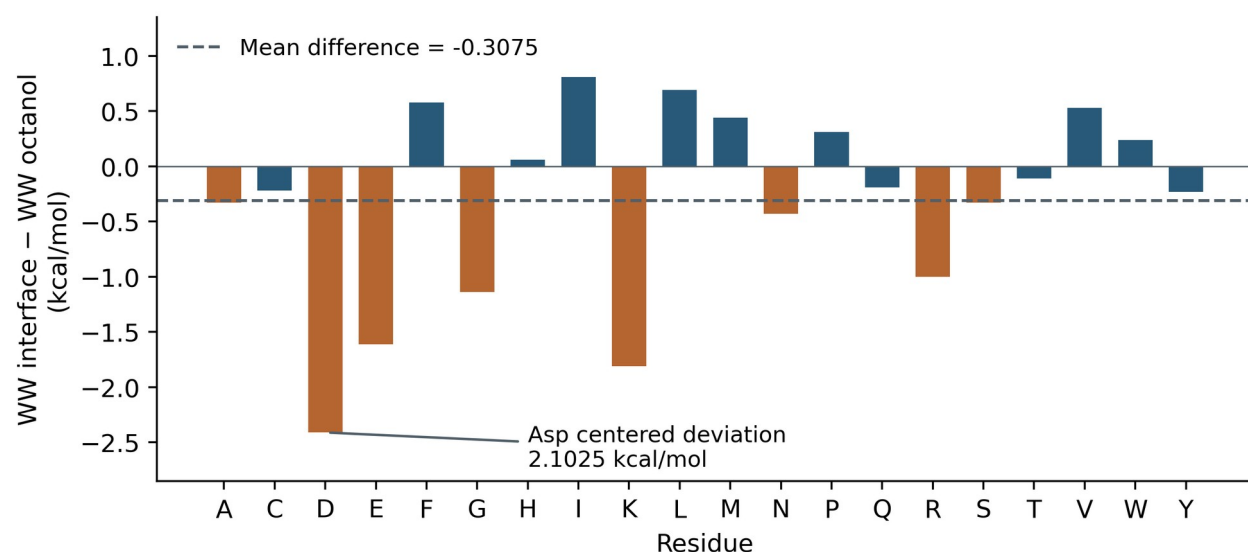


Figure 4. Residue-dependent environment interaction within the Wimley-White family. The per-residue difference WW interface minus WW octanol is not constant. The dashed line is the mean difference; orange and blue bars mark values below and above that mean, respectively. Aspartate has the largest absolute centered deviation (2.1025 kcal/mol), illustrating that strong raw correlation does not justify a residue-independent environment offset.

After the adversarial analysis, the thermodynamic mixed-contrast theorem survives exactly; q2 remains finite-slice exact; additive Born remains a state- and environment-conditional first-order model; conserved-support geometry survives its exact mathematical adversaries but has open physical bridges; and the unique microscopic mechanism remains nonidentifiable.

414 5.4 Prospective validation

415 No previously viewed endpoint is relabeled prospective. WW interface, Hessa, Moon-Fleming, Fauchere-Pliska,
416 Eisenberg, and the historical Radzicka-Wolfenden analogue benchmark remain retrospective/context evidence.

417 The primary no-refit prospective predictor was fixed before any new blind response data as

$$418 \quad \widehat{\Delta G} = 0.771447 - 0.357128 n_{CS} + 0.372280 n_{ON} + 16.778421 b_{Born}. \quad (16)$$

419 The preferred test is a new same-observer WW-type water-to-octanol panel of at least 12 noncanonical guest
420 residues, constructed from structures and state/protocol metadata before response values exist. **PASS** requires
421 external $Q^2 > 0$ and an exact one-sided Spearman permutation $p \leq 0.01$; **FAIL** requires $Q^2 \leq 0$ or $p > 0.05$;
422 intermediate outcomes are **INCONCLUSIVE**. No affine recalibration is allowed.

423 A counts-only no-refit comparator is separately fixed for a prospective Born-ablation gate: lower MAE and an
424 exact one-sided paired random-sign test at $p \leq 0.05$ are required. The existing contract already demands
425 chemical-family coverage; SI S16 now makes panel composition, assay precision, meaningful effect size and
426 design-specific power explicit premeasurement planning requirements. These requirements do not
427 retrospectively change the archived PASS rule. The guanidinium and independent $m = 4$ conserved-support tests
428 require robust rank certificates of two and three, respectively. A future crossed panel is also fixed to separate
429 Born-amplitude variation at fixed support count from support-count variation at approximately matched Born
430 amplitude.

431 At the present evidence state all genuinely new prospective tests remain **OPEN**. The archived formulas and
432 decision rules are reproducible; their existence does not establish independently timestamped preregistration,
433 and no prospective outcome is available.

434 5.5 Interpretation of the evidence

435 **What is exact.**

436 The exact physical contribution is a quotient statement. In a matched equilibrium phase-by-variant family, the
437 observer removes phase-only and variant-only free-energy gauges, leaving the mixed interaction component

438 $P_m F P_n$. This identifies what transfer double differences measure without assuming any hydrophobicity
439 descriptor. The validation-depth theorem is likewise exact: a predictive state is admissible only if its affine rank
440 survives every training domain reached by the complete validation protocol.

441 **What is finite-slice certified.**

442 The q2 state is exact on the declared canonical A20 WWO record and minimal within its five declared primitive
443 coordinates. Those are valuable certificates because they expose exactly which coarse physical distinctions fail
444 on that slice. They do not turn A20 into the full molecular state space, nor structural rank five into intrinsic
445 response dimension.

446 **What is empirical.**

447 The nested H1 results establish useful compression for WW octanol, WW interface, and Hessa under the
448 declared protocol, with qualified evidence for Moon-Fleming. Shared H2a prediction is strong enough to reject
449 random residue-label alignment in the declared campaigns. These results are predictive statements, not observer-
450 transfer or mechanism statements.

451 **What is not identified.**

452 Three nonidentifiability results are scientifically central. The scalar WWO endpoint does not identify latent
453 response rank. The current multi-endpoint design does not identify a unique shared geometry. The mechanism
454 competition leaves at least two distinct partial response-blind lifts with different falsifiers and open bridges;
455 therefore a unique complete microscopic mechanism is not identified.

456 **Why environment and observer belong inside the representation.**

457 The WW octanol/interface adversary demonstrates why context is not a nuisance variable to be erased after
458 fitting. Strong correlation can coexist with a large, residue-dependent environment interaction. Similar logic
459 governs charge-state handling: when an observer does not identify the state, the predictive grammar must lose
460 that distinction rather than invent it. Recent translocon, hydration, and sequence-context studies make the same
461 physical issue increasingly concrete: aqueous/interface/TM states can be experimentally distinguished [11],

hydration can change cooperatively with sequence patterning [12], and aromatic insertion propensity can depend strongly on the surrounding motif [14]. These results motivate typed observer contracts; they do not by themselves license pooling across observers. The observer determines which quotient is legal.

An exact response quotient is relative to its retained context family. Adding contexts can only refine that equivalence when all previous responses are retained; it does not establish coverage of unobserved environments. For an incomplete panel, standard graph-cycle algebra [36] identifies the observed interaction quotient with the cycle space of its bipartite observation graph, of dimension $|E| - m - n + c$, where c counts connected components. A tree contains no observed interaction contrast. These elementary extensions (SI S21) prevent apparent completeness from being created by zero filling or by treating unknown observations as physically inadmissible states.

Biophysical relevance without clinical promotion.

The immediate contribution is methodological and biophysical rather than clinical. By separating molecular state, environment, observer, exact interaction contrasts, and validation-safe predictive compression, the framework provides a reproducible basis for future studies of membrane insertion, mutation energetics, peptide/protein design, and physically constrained molecular modeling. Any translational or clinical use requires an independently validated domain adapter and is not established by the present amino-acid benchmark alone.

5.6 Limitations

The canonical A20 slice is small and chemically sparse, especially in charged families. More permutations improve resolution of the conditional null, not the number of independently observed molecules. The external structure-only demonstration establishes that the rank phenomenon occurs outside A20, but does not establish general predictive benefit across tasks. Several endpoint uncertainty ledgers are incomplete: Hessa lacks residue-wise uncertainty under the current source contract and Moon-Fleming row-level uncertainty provenance remains open. The physical q2 representation is not tautomer- or conformer-complete. Cross-endpoint observer transfer is partial or absent outside the WWO anchor. The $D \approx 3$ coefficient fingerprint is retrospective and

486 basis-specific. The mechanistic robust-rank and crossed-design tests have not yet produced new blind results.
487 These are not hidden caveats; they define the present claim boundary.

488 **6 Conclusion**

489 Physical sufficiency and predictive identifiability are distinct preservation requirements. The A20 physical state
490 q_2 is finitely sufficient yet fails nested-LOO rank safety because its sulfur distinction loses all training support
491 after two deletions. Within the declared D10 family, no nested-safe predictive class preserves the complete
492 physical state. Validation therefore constrains admissible unregularized representations before fitting, in
493 addition to estimating predictive error. Retrospective predictive signal, stable scores and shared profile
494 alignment do not identify unique physical coordinates or establish chemical extrapolation. Blind noncanonical
495 measurements are the next test of physical transferability.

496 **Reproducibility**

497 Supporting Information (Additional file 1) and the reproducibility archive (Additional file 2) contain the exact
498 E6 endpoint matrix, endpoint source/state records, the D10 descriptor ledger, the q_2 physical-state ledger, exact
499 interaction-quotient proof, exhaustive subset and deletion certificates, the cospark/nested-safety corollary
500 certificate, the finite-injective-encoding negative control, observer-transfer matrices, validation-safe state
501 counts, nested H1/H2 results, algorithm-matched permutation nulls, comparator and applicability analyses,
502 recovered historical coefficient and fixed-basis permutation replays, quantitative fiber-defect certificates,
503 singleton-Arg rank-loss controls, mechanistic boundary analyses, adversarial environment/chemistry records,
504 and declared prospective protocols. Machine-readable result files, executable replay scripts, environment
505 requirements, and a SHA-256 manifest permit independent integrity and numerical checks. The publication-
506 facing archive contains only materials needed to reproduce and audit the reported scientific results. The
507 verification scope is stated explicitly in the accompanying audit: integrity checks, exact structural
508 recomputation, numerical replay, and retained historical summaries are separate evidence categories.

509 **Additional files**

510 **Additional file 1. Supporting Information.** File:

511 **Additional_file_1_Supporting_Information_v1_0_6.pdf.** Format: PDF. Sections S1–S26 provide data
512 definitions and tables, proofs, detailed validation and permutation protocols, uncertainty and applicability
513 analyses, the claim ledger, literature comparisons, and the external structural audit. Figure S1 is included in this
514 file.

515 **Additional file 2. Reproducibility archive.** File:

516 **Additional_file_2_Reproducibility_Archive_v1_0_6.zip.** Format: ZIP. The archive supplies machine-
517 readable input data, source and state ledgers, analysis and verification code, numerical outputs, all figure
518 sources, software requirements, licenses, and SHA-256 integrity manifests. It supports independent reproduction
519 of the reported analyses without network access after the software dependencies have been installed.

520 **Abbreviations**

521 A20, canonical twenty-amino-acid panel; D10, ten-descriptor grammar; E6, six-endpoint panel; H1, endpoint-
522 specific prediction; H2a, shared prediction; H2b, shared-geometry identifiability; LOO, leave-one-out; MAE,
523 mean absolute error; OLS, ordinary least squares; PCR, principal component regression; PLS, partial least
524 squares; QSAR, quantitative structure-activity relationship; RMSE, root mean squared error; SSE, sum of
525 squared errors; TPSA, topological polar surface area; WW, Wimley-White; WWO, WW octanol; WWI, WW
526 interface.

527 **Declarations**

528 **Ethics approval and consent to participate.** Not applicable.

529 **Consent for publication.** Not applicable.

530 **Availability of data and materials.** The datasets supporting the conclusions of this article are included within
531 the article and its additional files. Additional file 1 contains the Supporting Information; Additional file 2

contains the exact endpoint and descriptor/state ledgers, machine-readable results, analysis and verification code, software requirements, and SHA-256 integrity manifests. All source code needed to reproduce the reported analyses is included in Additional file 2. Code is licensed under MIT; data, documentation, and figures under CC BY 4.0.

Competing interests. The author declares no competing interests.

Funding. No specific funding was received for this work.

Authors' contributions. AN conceived the study, designed the methodology, curated the data, implemented and verified the analyses, interpreted the results, and prepared the manuscript.

Acknowledgements. None.

Use of generative AI. Generative AI tools, including OpenAI ChatGPT, were used as research-assistance tools for code review, computational cross-checking, manuscript structuring, and language refinement. All hypotheses, protocol choices, data-provenance decisions, numerical results, interpretations, citations, and final manuscript content were reviewed and approved by the author, who assumes full responsibility for the work. No AI system is listed as an author.

References

1. Wimley WC, Creamer TP, White SH. Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. *Biochemistry*. 1996;35:5109-5124. doi:10.1021/bi9600153.
2. Wimley WC, White SH. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat Struct Biol*. 1996;3:842-848. doi:10.1038/nsb1096-842.
3. Hessa T, Kim H, Bihlmaier K, et al. Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature*. 2005;433:377-381. doi:10.1038/nature03216.
4. Moon CP, Fleming KG. Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers. *Proc Natl Acad Sci USA*. 2011;108:10174-10177. doi:10.1073/pnas.1103979108.

5. Eisenberg D, Schwarz E, Komaromy M, Wall R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J Mol Biol.* 1984;179:125-142. doi:10.1016/0022-2836(84)90309-7.
6. Fauchere JL, Pliska V. Hydrophobic parameters π of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides. *Eur J Med Chem.* 1983;18:369-375.
7. Radzicka A, Wolfenden R. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry.* 1988;27:1664-1670. doi:10.1021/bi00405a042.
8. Gu W, Rahi S, Helms V. Solvation free energies and transfer free energies for amino acids from hydrophobic solution to water solution from a very simple residue model. *J Phys Chem B.* 2004;108:5806-5814. doi:10.1021/jp0376424.
9. Breimann S, Frishman D. AAclust: k-optimized clustering for selecting redundancy-reduced sets of amino acid scales. *Bioinform Adv.* 2024;4:vbae165. doi:10.1093/bioadv/vbae165.
10. Breimann S, Kamp F, Steiner H, Frishman D. AAontology: An ontology of amino acid scales for interpretable machine learning. *J Mol Biol.* 2024;436:168717. doi:10.1016/j.jmb.2024.168717.
11. Grau B, Kormos R, Bano-Polo M, et al. Sequence-dependent scale for translocon-mediated insertion of interfacial helices in membranes. *Sci Adv.* 2025;11:eads6804. doi:10.1126/sciadv.ads6804.
12. Najafi S, Lobo S, Shell MS, Shea JE. Context dependency of hydrophobicity in intrinsically disordered proteins: insights from a new dewetting free energy-based hydrophobicity scale. *J Phys Chem B.* 2025;129:1904-1915. doi:10.1021/acs.jpcb.4c06399.
13. Ok Y, Park Y. One-dimensional molecular coordinate orders amino acids across water-mediated processes. *J Chem Inf Model.* 2026;66:10119-10126. doi:10.1021/acs.jcim.6c01897.
14. Muhammedkutty FNK, Zhou HX. A membrane insertion code for intrinsically disordered proteins. *eLife.* 2026;15:e111515. doi:10.7554/eLife.111515. Version of Record, 26 August 2026.
15. Hawkins DM, Basak SC, Mills D. Assessing model fit by cross-validation. *J Chem Inf Comput Sci.* 2003;43:579-586. doi:10.1021/ci025626i.
16. Baumann D, Baumann K. Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. *J Cheminform.* 2014;6:47. doi:10.1186/s13321-014-0047-1.

17. Ambure P, Gajewicz-Skretna A, Cordeiro MNDS, Roy K. New workflow for QSAR model development from small data sets. *J Chem Inf Model*. 2019;59:4070-4076. doi:10.1021/acs.jcim.9b00476.
18. Clark RD. A path to next-generation reproducibility in cheminformatics. *J Cheminform*. 2019;11:62. doi:10.1186/s13321-019-0385-0.
19. Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res*. 2010;11:2079-2107.
20. She Y, Tran H. On cross-validation for sparse reduced rank regression. *J R Stat Soc Series B Stat Methodol*. 2019;81:145-161. doi:10.1111/rssb.12295.
21. Phan TM, Phan TL, Van-Nguyen PC, et al. ProQSAR: A modular and reproducible framework for small-data QSAR modeling with fit-and-use models. *J Cheminform*. 2026;18:52. doi:10.1186/s13321-026-01175-9.
22. Waibl F, Fernandez-Quintero ML, Wedl FS, Kettenberger H, Georges G, Liedl KR. Comparison of hydrophobicity scales for predicting biophysical properties of antibodies. *Front Mol Biosci*. 2022;9:960194. doi:10.3389/fmolb.2022.960194.
23. Lobo S, Najafi S, Shell MS, Shea JE. Hydrophobicity in intrinsically disordered protein force fields: implications for conformational ensembles and protein–protein interactions. *J Phys Chem B*. 2025;129:6817-6827. doi:10.1021/acs.jpcb.5c02360.
24. Rekhi S, Mittal J. Amino acid transfer free energies reveal thermodynamic driving forces in biomolecular condensate formation. *Proc Natl Acad Sci USA*. 2025;122:e2425422122. doi:10.1073/pnas.2425422122.
25. Joeres R, Blumenthal DB, Kalinina OV. Data splitting to avoid information leakage with DataSAIL. *Nat Commun*. 2025;16:3337. doi:10.1038/s41467-025-58606-8.
26. Joeres R, Blumenthal DB, Kalinina OV. Addendum: Data splitting to avoid information leakage with DataSAIL. *Nat Commun*. 2026;17:1597. doi:10.1038/s41467-025-67495-w.

27. Guo Q, Hernandez-Hernandez S, Ballester PJ. Scaffold splits overestimate virtual screening performance. In: Artificial Neural Networks and Machine Learning – ICANN 2024. 2024:58-72. doi:10.1007/978-3-031-72359-9_5. Author version: arXiv:2406.00873.
28. Elnaggar A, Heinzinger M, Dallago C, et al. ProtTrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell. 2022;44:7112-7127. doi:10.1109/TPAMI.2021.3095381.
29. Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023;379:1123-1130. doi:10.1126/science.ade2574.
30. Notin P, Kollasch A, Ritter D, et al. ProteinGym: large-scale benchmarks for protein fitness prediction and design. Adv Neural Inf Process Syst. 2023;36. Official benchmark and publication record: <https://github.com/OATML-Markslab/ProteinGym>. Accessed 15 September 2026.
31. Pagano L, Toto A, Malagrino F, Visconti L, Jemth P, Gianni S. Double mutant cycles as a tool to address folding, binding, and allostery. Int J Mol Sci. 2021;22:828. doi:10.3390/ijms22020828.
32. Ertl P, Rohde B, Selzer P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. J Med Chem. 2000;43:3714-3717. doi:10.1021/jm000942e.
33. Zhong S, Zhao Y. Generic cospark of a matrix can be computed in polynomial time. In: 2017 IEEE International Symposium on Information Theory (ISIT). 2017:246-250. doi:10.1109/ISIT.2017.8006527.
34. Winkler AM, Webster MA, Brooks JC, Tracey I, Smith SM, Nichols TE. Non-parametric combination and related permutation tests for neuroimaging. Hum Brain Mapp. 2016;37:1486-1511. doi:10.1002/hbm.23115.
35. Mobley DL, Guthrie JP. FreeSolv: a database of experimental and calculated hydration free energies, with input files. J Comput Aided Mol Des. 2014;28:711-720. doi:10.1007/s10822-014-9747-x.
36. Jiang X, Lim LH, Yao Y, Ye Y. Statistical ranking and combinatorial Hodge theory. Math Program. 2011;127:203-244. doi:10.1007/s10107-010-0419-x.