Discovery and Validation of Physicochemical Determinants for Pan-Allele Peptide-MHC Class I Modelling Across Binding Affinity, Antigen Presentation, and Immunogenicity

by

Kevin Wang

Student ID: 1238288

A thesis submitted to Auckland University of Technology
in partial fulfilment of the requirements for the degree of
Master of Computer and Information Sciences, 2026.

Faculty of Design and Creative Technologies

Auckland University of Technology

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor material which to a substantial extent, has been accepted for the qualification of any other degree or diploma of a university or institution of higher learning, except when acknowledgement is made in the acknowledgements.

Signature: Date: 02/07/2026

Supervisor: Dr. Parma Nand

The work presented in this thesis was conducted under the supervision identified above. Where the work of others is used or cited, this is explicitly indicated in the text and in the references. External code dependencies are documented in Appendix A (software stack). The ESM2-35M backbone weights are released by Meta AI Research and are used under their published licence (`facebook/esm2_t12_35M_UR50D`). All other code, configuration, and analysis is original work of the author. The milestone vault `milestone_netmhcpan_beaten/` is released under the MIT licence.

Ethics Approval

This research did not require Auckland University of Technology Ethics Committee (AUTEC) approval. The work used only publicly available, de-identified datasets (NetMHCpan-4.1, IEDB, TransPHLA, TESLA, and BigMHC) and did not involve human participants, animals, or identifiable private information.

Acknowledgements

I am deeply grateful to my supervisor, Dr. Parma Nand, for steady guidance, critical feedback, and sustained encouragement throughout this research. I thank Auckland University of Technology and the Faculty of Design and Creative Technologies for academic and institutional support, and I acknowledge the colleagues and collaborators whose discussions helped refine the methodology and presentation.

The author acknowledges the use of AI tools for preliminary background research and assistance with document formatting; however, all core research, data analysis, and original findings presented in this thesis are the independent work of the author.

Abstract

Problem. Personalised cancer-vaccine design requires predicting which of a patient's mutated peptides will be presented to the immune system. The decision rests on three sequential biological filters: a peptide must bind a major histocompatibility complex (MHC) class I molecule (*binding affinity*, BA), the complex must be processed and displayed at the cell surface (*eluted-ligand presentation*, EL), and the displayed complex must be recognised by a $CD8^+$ T-cell receptor (*immunogenicity*, IM). Computational prediction of the first two filters has been the rate-limiting step for two decades, and NetMHCpan-4.1 remains the field's reference predictor.

Gaps and research questions. Reviewing that lineage identifies three gaps. Existing pan-allele predictors transfer poorly to the long tail of low-frequency human leukocyte antigen (HLA) alleles; they under-fit non-canonical peptide lengths; and they consume the 34-residue MHC *pseudo-sequence* — the structurally derived map of which binding-groove positions contact the peptide — only as a flat embedding or an allele identifier, never as a structural target that individual peptide positions can address. A fourth, methodological concern is that headline comparisons in this literature are often reported without confidence intervals, from single folds, or without decontaminating the benchmark against the dominant training corpus. These motivate four research questions: whether position-addressable cross-attention over the pseudo-sequence improves prediction (RQ1); whether contrastive regularisation grouped by structural similarity rather than allele identity improves rare-allele transfer (RQ2); whether one pan-length model suffices for non-canonical lengths (RQ3); and whether, under a declared evidential contract, the resulting predictor outperforms NetMHCpan-4.1 — and where it does not (RQ4).

Approach. This thesis develops **DeepNeo-CL**, a peptide–MHC-class-I predictor built on a frozen ESM2-35M protein language-model trunk extended with position-resolved physicochemical features, peptide-conditioned cross-attention over the pseudo-sequence, joint BA and EL output heads, long-tail imbalance correction, and a supervised contrastive refinement grouped by pseudo-sequence cluster. A compact immunogenicity head reuses the same trunk, so one forward pass returns BA, EL and IM scores. Training used the official NetMHCpan-4.1 five-fold corpus (207,029 BA pairs and approximately 33 M EL pairs), and every Stage 1 result reported here is a five-fold ensemble. The Stage 2 immunogenicity head is trained and reported separately, as a three-seed ensemble over the frozen Stage 1 trunk.

Experiments. Evaluation used three independent held-out benchmarks — TransPHLA ($n = 161{,}652$), an IEDB BA snapshot decontaminated at peptide–HLA pair level against the training corpus ($n = 1{,}365$), and two years of the official IEDB weekly `auto_bench` series — plus BigMHC's held-out immunogenicity test set for the Stage 2 head. Comparators were NetMHCpan-4.1 and MHCflurry-2.0, extended to BigMHC, MHCfovea, TransPHLA and ImmuneApp on identical post-cutoff rows. A declared three-condition evidential contract — bootstrap confidence

intervals, paired DeLong testing, and per-allele Wilcoxon agreement — is the standard against which the comparisons are reported, and the thesis states plainly what that standard is worth: it is not a pre-registration, and it was not applied in full. The DeLong condition is defined for the area under the receiver-operating-characteristic curve (AUROC) only, so no comparison of areas under the precision–recall curve (AUPR) lies within its reach; the paired bootstrap interval the first condition asks for was never computed, so no verdict satisfies all three. A cumulative ablation isolated each component's contribution.

Results. The outcome is two-sided. On TransPHLA the ensemble exceeds NetMHCpan-4.1 on three of four metrics and ties on the fourth: BA AUROC/AUPR reach $0.9694/0.9705$ through a label-free BA+EL rank fusion (DeLong $z = 22.9$), EL AUROC reaches 0.9697 ($p = 1.8 \times 10^{-8}$), and EL AUPR is tied (0.9714 vs 0.9716). On IEDB the two predictors are *not distinguished* on within-allele ranking across two years of weekly data (Wilcoxon $p = 0.65$ — a failure to detect a difference, not a demonstration of equivalence), while on the static snapshot DeepNeo-CL *trails* both on pooled cross-allele ranking and within allele, and percentile-rank recalibration does not repair that deficit. MHCflurry-2.0 is tied on IEDB. The ablation shows no single component worth more than $+0.005$ AUROC. Taken in pairs the contributions are larger but still modest: the pseudo-sequence family — concatenation together with cross-attention — contributes $+0.007$ after the physicochemical rung, the contrastive refinement is within noise pooled yet worth $+0.022$ AUPR on rare alleles, and the fusion lifts BA AUROC from the BA head's 0.9546. An exploratory structural rescorer was *withdrawn* when its advantage proved transductive. The immunogenicity head reaches AUROC 0.5735, a statistical tie with BigMHC's published predictor ($p = 0.70$), with no dedicated immunogenicity predictor separating from it after multiple-comparison correction.

What this means. The answer to RQ1 is a qualified yes — the mechanism works, but its isolated effect is small and it costs a fixed ≈ 10 per cent of the forward pass, so it is a modest trade rather than a free one. RQ2 is answered yes on the subgroup it targets, and the structural grouping rather than contrastive learning as such is what carries it. RQ3 is answered only partly, since the benchmark covers 8–11mers and cannot speak to the longer peptides that motivated the gap. RQ4 is answered three ways: yes on TransPHLA, tied on IEDB within-allele, and no on pooled IEDB. The defensible conclusion is therefore narrow: several biologically motivated refinements can together produce a measurable improvement on an independent benchmark, with the margins small, one benchmark unfavourable, and the evidential standard declared in advance of the results but neither pre-registered nor fully applied. The within-allele tie and the pooled deficit are measured on different corpora — two years of auto_bench and a single 1,365-row snapshot respectively — so together they bound what can be claimed rather than isolating a single cause; percentile-rank recalibration does not repair the pooled deficit, so it is not a score-scale artefact. The thesis makes no claim on TCR specificity, class II MHC, or post-translationally modified peptides, and advances no clinical claim. Alongside the thesis a **reproducibility vault** is released — code, ensemble checkpoints with integrity hashes, raw prediction arrays, statistical logs, figures, and a CPU-only validator — from which the base-ensemble results can be re-derived. The thesis also records which results lie outside it, the fused headline figure among them, and names the artefacts that produce those, so that every reported number can be traced to its source even where that source is not the vault.

Contents

Attestation of Authorship i

Ethics Approval ii

Acknowledgements iii

Abstract iv

List of Equations xxiii

List of Abbreviations xxiv

Notation xxv

Author Statement of Responsibility xxvi

1 Introduction 1

1.1 Motivation . . . 1

1.1.1 Stage taxonomy at a glance . . . 1

1.2 The state of pMHC-I prediction in 2026 . . . 2

1.3 Three gaps in current pan-allele predictors . . . 3

1.4 Research questions . . . 4

1.5 Thesis statement and scope . . . 5

1.6 Contributions . . . 5

1.7 Roadmap . . . 6

2 Literature review and background 8

2.1 Scope of the review . . . 8

2.2 Evolution of the field . . . 9

2.2.1 Position-specific scoring matrices (1995–2007) . . . 9

2.2.2 Pan-allele neural era (2007–2020) . . . 9

2.2.3 MHCflurry-2.0 . . . 9

2.2.4 Transformer era (2022–) . . . 9

2.2.5 Foundation-model era . . . 9

2.2.6 The lineage assessed: mechanism, contribution, limitation, suitability . . . 10

2.3 Architectural devices that have worked . . . 12

2.3.1 Pseudo-sequence concatenation . . . 12

2.3.2 Multi-task BA + EL joint training . . . 12

2.3.3 Position-resolved physicochemical features . . . 12
2.3.4 Imbalance correction . . . 12
2.3.5 Contrastive regularisation . . . 12
2.3.6 Structural priors at inference . . . 12
2.4 Methodological gap this thesis fills . . . 13
2.5 What this thesis explicitly is *not* claiming . . . 13
2.6 Summary of the literature . . . 14
2.7 Orientation: what a computing reader needs from this biology . . . 14
2.8 The MHC class I complex . . . 15
2.9 Anchor positions . . . 15
2.10 The proteomic pipeline . . . 16
2.11 Allele polymorphism and the long tail . . . 16
2.12 The neoantigen problem . . . 17
2.13 Why structure matters . . . 17
2.14 Chapter summary . . . 18

3 Methodology: datasets, materials and methods 19
3.1 Research design . . . 19
3.2 Materials . . . 20
3.3 Overview . . . 20
3.4 Training corpus: NetMHCpan-4.1 official 5-fold . . . 21
3.4.1 Stratification invariants . . . 21
3.4.2 Why upstream byte-for-byte . . . 21
3.5 Independent benchmark: IEDB BA . . . 21
3.5.1 Source . . . 21
3.5.2 Decontamination protocol . . . 21
3.5.3 Final size . . . 22
3.5.4 Statistical power . . . 22
3.6 Independent benchmark: TransPHLA . . . 23
3.6.1 Source . . . 23
3.6.2 Statistical power . . . 23
3.6.3 Co-primary role . . . 23
3.7 Zero-shot benchmark: TESLA . . . 23
3.8 Pocket-feature caches (withdrawn structural experiment) . . . 23
3.9 Homology-reduction sensitivity analysis . . . 24
3.10 Data lifecycle diagram . . . 24
3.11 Reproducibility and provenance . . . 27
3.12 Summary of the datasets . . . 27
3.13 Methods . . . 28
3.14 Notation . . . 28
3.15 Backbone: ESM2-35M . . . 30
3.16 Peptide × HLA pseudo-sequence cross-attention . . . 31
3.17 Multi-task BA + EL heads . . . 31
3.17.1 Binding-affinity head: censored MSE . . . 31

3.17.2 Presentation head: focal BCE . . . 32
3.17.3 Joint objective . . . 32
3.18 Imbalance correction (E11 + E12 → E13) . . . 32
3.18.1 Label-distribution smoothing (LDS) . . . 32
3.18.2 Effective-number weighting on pseudo-cluster × bin . . . 33
3.18.3 Combined Stage 1 objective . . . 33
3.19 SupCon refinement experiment: supervised contrastive regularisation (SupCon v2) 34
3.20 Structural pocket rescorer (explored, withdrawn) . . . 36
3.20.1 The 370-D crystal-template pocket descriptor . . . 36
3.20.2 The residual reranker MLP . . . 36
3.20.3 Cross-validation . . . 36
3.21 Inference pipeline . . . 37
3.22 Computational budget summary . . . 37
3.23 5.9b Software, hardware, and what the budget decided . . . 37
3.23.1 5.9b.1 What reproducibility this work delivers, and what it does not . . . 38
3.24 Implementation and its reproducibility boundary . . . 38
3.25 Chapter summary . . . 39

4 Model architecture **40**
4.1 Overview diagrams . . . 40
4.2 Component inventory . . . 43
4.2.1 The forward pass, component by component . . . 43
4.3 Hyperparameter justifications . . . 45
4.4 What we did *not* include in the final architecture . . . 45
4.5 Comparison with reference architectures . . . 47
4.6 Computational complexity, memory, and scalability . . . 47
4.6.1 Asymptotic cost of a forward pass . . . 47
4.6.2 Memory footprint . . . 49
4.6.3 Measured throughput and scaling limits . . . 49
4.6.4 Parameter efficiency against the reference predictors . . . 49
4.7 Summary . . . 51

5 Experimental design **52**
5.1 The experiments, and which question each answers . . . 52
5.2 Metric definitions . . . 54
5.2.1 Area under the receiver-operating-characteristic (ROC) curve . . . 55
5.2.2 Area under the Precision–Recall curve . . . 55
5.2.3 Spearman rank correlation . . . 56
5.2.4 PPV at the top-k% . . . 56
5.3 Bootstrap confidence intervals . . . 56
5.4 DeLong paired-AUROC test . . . 57
5.5 Per-allele Wilcoxon signed-rank consistency . . . 58
5.6 The three-condition contract . . . 59
5.6.1 What the contract is, and is not, evidence of . . . 59
5.6.2 What the contract can and cannot adjudicate . . . 60

5.7 Eight-rule fair-comparison protocol . . . 60
5.8 Ensemble protocol . . . 62
5.9 What we explicitly do NOT do . . . 62
5.10 Reproducibility envelope . . . 63
5.11 Summary . . . 64

6 Results **66**
6.1 Scope of reported results . . . 67
6.2 Headline table . . . 67
6.3 Per-length breakdown (TransPHLA EL) . . . 69
6.4 Per-allele forest plot (IEDB BA) . . . 70
6.5 Deployment-safe BA+EL fusion . . . 73
6.5.1 The fusion recipe . . . 73
6.5.2 TransPHLA: where fusion helps (and the symmetric analysis) . . . 73
6.5.3 IEDB: fusion improves within-allele ranking . . . 73
6.6 Common vs rare allele transfer . . . 76
6.7 Operational PPV at top-k% . . . 78
6.8 MHCflurry-2.0 comparison (secondary) . . . 79
6.9 External head-to-head against post-2020 predictors . . . 79
6.10 Statistical-significance summary . . . 79
6.11 Homology-reduction sensitivity (IEDB BA) . . . 81
6.12 Summary . . . 81

7 Ablation studies **83**
7.1 Cumulative ablation grid . . . 83
7.2 The E13 imbalance-correction saga (E5–E14) . . . 85
7.2.1 E5 — Pocket-anchor loss bias . . . 86
7.2.2 E6 — Allele-hard sampler . . . 86
7.2.3 E7 — Hybrid BA sampler (E5 + E6 combination) . . . 86
7.2.4 E8, E9 — Strong-binder tail weighting (diagnostic) . . . 87
7.2.5 E10 — Adaptive length × BA-bin . . . 87
7.2.6 E11 — LDS-weighted BA regression . . . 87
7.2.7 E12 — Effective-number BA subgroup weighting . . . 87
7.2.8 E13 — Joint (`pseudo_cluster` × `ba_bin`) effective-number (THE WINNER) . . . 87
7.2.9 E14 — Model-weight EMA . . . 87
7.2.10 What the ten attempts establish . . . 88
7.3 Cross-attention (xattn) . . . 89
7.4 SupCon contrastive ablations . . . 89
7.4.1 Phase 2 SupCon v1 — clean configuration (ba_bin label, $\tau = 0.07$, w = 0.3) 89
7.4.2 Phase 2 SupCon v2 — pseudo_cluster label, fp32, $\tau = 0.2$, w = 0.1 . . . 89
7.4.3 Phase 2 SupCon v3 — `hla_id` label . . . 89
7.5 Pocket-reranker ablations . . . 90
7.5.1 Phase 3 v1 — ESMFold structural features . . . 90
7.5.2 Phase 3 v2 — crystal-template pocket descriptor (WITHDRAWN) . . . 91

7.6 Phase 1 ablations (deferred) . . . 91
7.6.1 Phase 1 — masked-language-model continued pretraining, and full fine-tuning as the LoRA-equivalent . . . 91
7.7 Methodology audit ablations . . . 91
7.7.1 Audit fix #1 — Proper E13 5-fold ensemble . . . 91
7.7.2 Audit fix #2 — Proper SupCon v2 5-fold ensemble . . . 92
7.8 Summary of dispositions . . . 92
7.8.1 Criteria for each verdict . . . 92
7.9 Why the ablation pattern matters . . . 93
7.10 Summary . . . 95

8 Discussion **97**
8.1 Why DeepNeo-CL separates from NetMHCpan-4.1 on TransPHLA . . . 98
8.2 Long-tail allele transfer . . . 98
8.3 Calibration . . . 99
8.4 Where DeepNeo-CL ties on IEDB, and where it does not . . . 100
8.4.1 Sample size . . . 100
8.4.2 Allele mix . . . 100
8.4.3 Label noise . . . 101
8.4.4 Capacity ceiling at ESM2-35M . . . 101
8.5 Comparison with MHCflurry-2.0 . . . 101
8.6 What the win is *not* . . . 102
8.7 The methodological contribution . . . 102
8.8 What the result says about pMHC-I biology . . . 103
8.9 Summary . . . 103

9 Limitations **105**
9.1 IEDB BA is tied, not won . . . 105
9.2 Capacity ceiling at ESM2-35M . . . 106
9.3 Single-fold results are not the headline . . . 106
9.4 The pocket reranker is withdrawn (transductive) . . . 106
9.5 MHCflurry-2.0 EL is not benchmarked . . . 107
9.6 Limited Stage 2B scope; no clinical claim . . . 107
9.7 Class II is out of scope . . . 107
9.8 Post-translational modifications . . . 107
9.9 Schematic / partial figures . . . 107
9.10 Reproducibility of the *training* run . . . 108
9.11 Software stack . . . 108
9.12 Three statistical exposures, with direction and, where computable, magnitude . . . 108
9.13 Summary . . . 111

10 Conclusion and future work **112**
10.1 Answers to the research questions . . . 112
10.2 Summary of contributions . . . 114
10.3 The single defensible claim, as discharged . . . 115

10.4 What the milestone vault enables . . . 115
10.5 Honest reflection on what this thesis is not . . . 116
10.6 What this thesis claims for the field . . . 116
10.7 The Stage 1 -> Stage 2 bridge . . . 116
10.8 Where the two stages leave the pipeline . . . 117
10.9 Stage 2A TCR-conditioned modelling (highest-ROI) . . . 118
10.9.1 Proposed protocol . . . 118
10.10 ESM2-150M Stage 1 redux (deferred, low risk) . . . 118
10.10.1 Rationale . . . 118
10.10.2 Risk-cost calculus . . . 118
10.11 Class II MHC extension . . . 119
10.12 Peptide:MHC:TCR triplet modelling (Stage 2A) . . . 119
10.13 Distillation for clinical-pipeline throughput . . . 120
10.14 AlphaFold3-API features for a next structural-reranker experiment . . . 120
10.15 Adversarial robustness . . . 120
10.16 Vault-as-template: methodological adoption . . . 121
10.17 Engineering and tooling . . . 121
10.18 Analyses that would strengthen the present claims . . . 122
10.18.1 Leave-one-out ablation to complement the cumulative ladder . . . 122
10.18.2 Per-pocket gradient attribution over the trained model . . . 122
10.18.3 Unfrozen-trunk transfer learning as a comparator . . . 123
10.18.4 The homology-reduction sensitivity analysis, which was specified and not run 123
10.19 Summary of priorities . . . 124

A Reproducibility appendix **125**
A.1 Layered reproducibility model . . . 125
A.2 The milestone vault (`milestone_netmhcpan_beaten/`) . . . 125
A.3 L1 — Read-only reproduction . . . 126
A.4 L2 — Integrity check (< 2 minutes CPU) . . . 126
A.5 L3 — Stats re-derivation (~ 15 minutes on A100) . . . 127
A.6 L4 — Full from-scratch retrain (~ 17.5 GPU-hours) . . . 127
A.6.1 Environment . . . 127
A.6.2 Stage 0 — Data preparation . . . 127
A.6.3 Stage 1 — E13 5-fold . . . 128
A.6.4 SupCon refinement — v2 continuation . . . 128
A.6.5 Structural pocket rescorer (withdrawn experiment) . . . 128
A.6.6 Evaluation . . . 129
A.6.7 Vault rebuild . . . 129
A.7 Expected variance . . . 129
A.8 RunPod environment specifics . . . 129
A.9 Software-stack versions (frozen 2026-05-20) . . . 130
A.10 How to cite this vault . . . 130
A.11 Reproducibility checklist [1] . . . 130
A.12 Implementation inventory . . . 131

A.13 Architecture-metadata snapshot . . . 132
A.14 Training dynamics . . . 132

B Stage 2B: T-cell-recognition prediction (introduction) 134
B.1 Why Stage 2: extending the BA → EL → IM pathway end-to-end . . . 134
B.2 What Stage 2 ships . . . 135
B.3 Pipeline overview . . . 135
B.4 Stage 2B contributions . . . 135
B.5 Stage 2A — what we do not claim . . . 135
B.6 Scope of reported results . . . 137
B.7 Appendix roadmap . . . 137
B.8 End-to-end formulation: BA → EL → IM on one shared trunk . . . 137
B.9 Stage 2B immunogenicity head architecture . . . 138
B.10 Loss function . . . 138
B.11 Training protocol . . . 138
B.12 Statistical contract (carries from Stage 1) . . . 138
B.13 Scope statements carried into the sections below . . . 139
B.14 Computational budget . . . 139
B.15 Why not full / LoRA fine-tuning? . . . 140
B.16 Reproducibility . . . 140
B.17 Benchmarks . . . 140
B.18 Reference predictors (head-to-head) . . . 141
B.19 Metrics . . . 141
B.20 Significance tests . . . 141
B.21 The eight-rule fair-comparison protocol (carries from Stage 1) . . . 141
B.22 Honest-statistical expectation (from literature) . . . 142
B.23 Acceptance criteria for the Stage 2B run . . . 142
B.24 Reference benchmarks for Stage 2A (out of scope) . . . 142
B.25 Scope of inference . . . 142
B.26 Canonical headline result (E10, 3-seed ensemble, frozen trunk) . . . 143
B.27 Statistical interpretation . . . 143
B.28 Legacy audit trail (non-headline) . . . 144
B.29 Benchmark difficulty and ceiling effects . . . 144
B.30 Secondary benchmark context (TESLA) . . . 144
B.31 Chapter summary . . . 144
B.32 Supplementary Stage 2B result panels . . . 144
B.33 Main interpretation of Stage 2B . . . 148
B.34 Why parity and residual gap coexist . . . 148
B.35 Benchmark difficulty and ceiling behavior . . . 148
B.36 Biological implications . . . 148
B.37 Methodological implications . . . 148
B.38 Scope boundaries . . . 149
B.39 Link to next-stage development . . . 149
B.40 Stage 2B in one paragraph . . . 149

B.41 Stage-2 contribution to the thesis . . . 149
B.42 What remains unresolved . . . 150
B.43 Next technical step: Stage 2A . . . 150
B.44 Final Stage-2 claim statement . . . 150
B.45 Closing . . . 150

C Project Documentation Index **151**
C.1 Canonical stage taxonomy . . . 151
C.2 Controlled documentation set . . . 151
C.3 Reproducibility anchors . . . 151

D Longitudinal IEDB validation on the official weekly benchmark (2024–2026) **153**
D.1 Motivation . . . 153
D.2 Data and evaluation protocol . . . 153
D.3 The benchmark metric: official IEDB scoring versus what we report . . . 154
D.4 Results . . . 155
D.4.1 Pooled two-year standing . . . 155
D.4.2 Deployment-fair verdict: a statistical tie . . . 155
D.4.3 Per-allele tendency . . . 156
D.5 Head-to-head against post-2020 deep predictors . . . 157
D.5.1 Data preparation . . . 158
D.5.2 Method versions and provenance . . . 158
D.5.3 Deployment-fair verdict at 2020–2024 scale . . . 159
D.5.4 Architectural interpretation . . . 160
D.5.5 Threats to validity for this section . . . 162
D.6 Limitations . . . 163
D.7 Claim boundary and reproducibility . . . 163

List of Figures

3.1 End-to-end data lifecycle, from raw sources to benchmark-ready assets. **Every corpus here is upstream**: what this thesis contributes to the data layer is the middle column — the decontamination protocol, the allele-availability filter, and the evidential contract that follows from them — and the figure is shaded to make that distinction rather than leaving it to be inferred. The dashed edge is not data flow: the training corpus is the *reference set* the decontamination is measured against, which is a different relationship from the rest of the diagram. The three caveats are placed in the figure rather than only in the caption, because a lifecycle diagram that showed only the path on which everything worked would be the kind of figure Chapter 9 exists to argue against: the decontamination is **exact-pair only**, the homology pass was specified and never run (§3.9), and the registry records metadata sidecars rather than MD5 manifests (§3.5.2, §3.11). 25

3.2 The four pipelines and the software stack beneath them. Every path and script name is a real file in the project repository. Two properties of the design are load-bearing and are marked: a checkpoint is resolved by **MD5**, never by filename — two checkpoints in this project carry the same recipe name and differ in both output heads, so name resolution would silently select the wrong model — and every reported number resolves to a committed artefact rather than to a figure carried forward from an earlier draft. No performance figure appears here either; this is how results are produced, not what they were. 26

3.3 The research process, in the order it was carried out, with the research question each stage serves. Solid arrows are sequence. The dashed arrow marks governance rather than data flow: the evaluation contract of §5.6 is the declared standard the comparisons at steps 7–9 are reported against. **It is not a pre-registration** — it was written down after the first ensemble results were in hand, and §5.6.1 states exactly what it is and is not worth. The diagram places it at step 4 because that is where it sits in the logical order of the argument, not because it was authored there. Step 9 is a withdrawn branch, retained for the reason given in the text. No performance figure appears here; the results are in Chapter 6. 29

4.1 DeepNeo-CL Stage 1 at the level of information flow. Reading left to right: the **peptide branch** embeds the 8–11-residue peptide with the shared ESM2-35M encoder; the **pseudo-sequence branch** passes the 34-residue HLA pocket string through the *same* encoder, which is what makes the two representations commensurable; the **cross-attention block** lets peptide positions query pocket positions, so allele conditioning is position-addressable rather than a single learned vector; the **physicochemical branch** supplies per-residue V37 descriptors that the language model does not carry; and the **fusion trunk** concatenates these into one representation read by the two task heads. This figure shows *what is connected to what*; for layer counts, tensor dimensions and parameter counts see Figure 4.3. . 41

4.2 The same Stage 1 model drawn for implementation reading, with the training-time components that Figure 4.1 omits. The **shared trunk** is the ESM2-35M encoder — frozen except for its last two layers — plus the fusion layers; the **BA head** regresses a censored affinity target and the **EL head** classifies presentation, both from the same representation, which is the sense in which this is one model rather than two; the **SupCon refinement** is a four-epoch continuation that reshapes that representation by pulling together pairs from the same pseudo-sequence cluster. SupCon is a *training-time* objective only — it adds no parameters and no inference cost, and is absent at prediction time. 42

4.3 The Stage 1 network as built, at layer and tensor granularity. Every parameter count and tensor dimension is read from the released fold-`c000` checkpoint rather than from the design description, by `thesis_final/scripts/dump_architecture.py`; the component figures sum exactly to the checkpoint's 36,046,593 parameters, and the script fails if they stop doing so. Two values shown here are *not* recoverable from the weights and come from the code and the configuration instead: the attention head count — `nn.MultiheadAttention` stores one $(3d, d)$ projection whatever the head count — and the dropout rate. The artefact records both as such. Three details are visible here and in neither diagram above: the cross-attention block's output projection is *zero-initialised*, so it begins as an identity on the peptide tokens and learns to add HLA context rather than starting from noise; the supervised-contrastive loss is applied directly to the shared 256-D representation, with no projection head of the kind the contrastive literature normally places there; and the BA and EL heads are single linear read-outs of that same representation, which is what makes them two views of one encoder rather than two models. . . 44

6.1 ROC and PR curves for deployment-safe headline comparisons (TransPHLA BA panel uses BA+EL rank-average fusion; IEDB and TransPHLA EL use the base 5-fold ensemble). 68

6.2 Per-length decomposition of TransPHLA EL performance (8–11mers only; the benchmark contains no 12–14mers). 69

6.3 Per-allele AUROC deltas on IEDB BA (DeepNeo-CL minus NetMHCpan-4.1), with **paired** 95% DeLong confidence intervals on the difference. HLA-A*33:03 is omitted because it is single-class, so its AUROC is undefined (§3.5.3), leaving ten alleles. Only HLA-A*02:01 excludes zero under both estimators tried; HLA-A*68:01 ($n = 38$) is estimator-sensitive and is discussed in §6.4. The interval is paired because the single-predictor Hanley–McNeil standard error would omit the comparator's variance and the pairing correlation. 71
6.4 Common-versus-rare allele transfer behavior on IEDB BA. 77

7.1 Cumulative ablation ladder (TransPHLA EL, 5-fold ensemble) showing the incremental contribution of each component. The table in §7.1 is the deployment ladder and stops at SupCon v2; the bottom *hatched* rung (pocket reranker) is withdrawn-transductive context only, not a deployment result. 84
7.2 SupCon embedding-space diagnostic t-SNE view (illustrative). 90

8.1 Calibration reliability comparison on TransPHLA EL. 99

A.1 Representative Stage 1 (BA/EL) training dynamics for fold c000. Curves are schematic and illustrate the qualitative training behavior discussed in the text. . . 132

B.1 Stage 2B architecture: frozen Stage 1 trunk, 2-layer immunogenicity head, 3-seed ensemble, and paired statistical evaluation contract. 136
B.2 Canonical Stage 2B frozen-trunk E10 3-seed head-to-head on BigMHC `im_test`. 144

D.1 Per-dataset paired AUROC, DeepNeo-CL versus NetMHCpan-4.1 156
D.2 Per-week AUROC, DeepNeo-CL versus NetMHCpan-4.1 157
D.3 Head-to-head vs post-2020 deep predictors 161

List of Tables

1.1 Canonical stage taxonomy used throughout the thesis. The right-hand column is resolved by cross-reference rather than written out, so it cannot fall out of step with the chapter numbering again. 2

2.1 The pMHC-I prediction lineage assessed. "Usable here" is the operative column: a method is usable if this thesis can either build on its mechanism or compare against it on identical rows. Each cell records the primary use; the pan-allele shallow nets are the one generation the thesis does both with — NetMHCpan-4.1 supplies the pseudo-sequence device (§2.3.1) and is the reference comparator — while the PSSMs supply motif evidence whose scoring model is superseded, and the language models supply a trunk that cannot itself be compared against. Note that from the pan-allele generation onward every row that describes an allele does so by a sequence of groove residues rather than an identifier; the protein-language-model row describes no allele at all, its trunk being allele-agnostic. 11

3.1 Corpora used in this thesis. The NetMHCpan-4.1 official five-fold corpus supplies all training data; IEDB BA and TransPHLA are held-out independent benchmarks; and TESLA is a zero-shot positive control, not a source of reported immunogenicity numbers (§3.7). Row counts are the figures actually scored, and only one of them is a decontaminated count: IEDB BA's 1,365 is *after* pair-level decontamination against the training folds, reduced from 4,800 (§3.5.3), while TransPHLA and TESLA are used at their published extent for the reasons §3.6.1 gives. 20

3.2 Structural pocket-feature caches, retained for provenance only. These were produced for the crystal-template rescoring experiment (§3.20), which is withdrawn as transductive; no deployment result or headline number in this thesis depends on them. 24

3.3 Wall-clock computational budget per pipeline stage, measured on a single NVIDIA A100 40 GB. Per-fold figures multiply by five for the complete ensemble. Chapter 4 §4.6 gives the asymptotic cost of the architecture and the measured inference throughput. 37

4.1 Component inventory of the Stage 1 DeepNeo-CL model, with the parameter count and functional role of each block. Parameter counts are for a single fold; the deployed predictor is a five-fold ensemble of this architecture. Counted from the released checkpoint, the components total 36,046,593 parameters, of which the ESM2-35M trunk accounts for 33,500,413 and 2,546,180 sit outside it. **Those** 2,546,180 **are not the trainable count.** The trunk is frozen *except* for its last two transformer layers (`finetune_last_n_layers: 2`), which contribute a further 5,542,080, so **8**,**088**,**260** parameters — 22.4 % of the model, not 7.1 % — receive gradient. The rows below are component totals and do not partition by what is frozen. 43

4.2 Non-default hyperparameter choices and the ablation evidence behind each. Every value *in this table* traces to a documented experiment in the Stage 1 ablation log; the final column gives that reference. Choices left at their published defaults are not listed, and two non-default values that no experiment selected — batch size and fusion trunk width — are named in the chapter introduction rather than here. . . . 46

4.3 Architectural comparison against the reference predictors, on characteristics rather than accuracy: representational backbone, how the allele is encoded, whether binding affinity and eluted-ligand presentation are learned jointly, whether a post-hoc rescoring stage is used, and parameter count. Accuracy comparisons on matched rows are reported separately in Chapter 6. 47

5.1 Every experiment reported in this thesis, with the question it serves and where its result appears. "Negative" marks an experiment whose outcome did not support the component it tested; "not run" marks one that was specified and never executed. Each is reported in full at the section given in the last column. The pocket rescorer is a distinct case and is not marked negative: it *passed* this chapter's contract and was withdrawn on a ground the contract does not contain, which §5.11 sets out. . 53

5.2 The three-condition statistical contract. A difference between DeepNeo-CL and a reference predictor is reported as real only when all three conditions hold simultaneously; any single condition failing downgrades the result to a tie, however favourable the point estimate. Its provenance is stated below: the standard it elaborates was fixed on 2026-05-17, this particular conjunction was written down on 2026-05-18, and §5.6.2 records where it was not met. 59

6.1 Headline comparison on the two independent benchmarks, scored deployment-safe with no evaluation-set fitting. Bracketed figures are bootstrap 95 % confidence intervals. Reading the table: on TransPHLA DeepNeo-CL is higher on BA AUROC, BA AUPR and EL AUROC and tied on EL AUPR, with the BA rows obtained through the label-free BA+EL rank fusion of §6.5; the deltas are small (+0.0015 to +0.0063) but resolvable at $n = 161{,}652$. On the 1,365-row IEDB snapshot the pooled figures favour NetMHCpan-4.1 by a margin that *is* resolvable at that sample size — −0.0651 AUPR is roughly 3.3 standard errors (§3.5.4) — so this is a measured deficit and not a limit of n. The within-allele analysis of §6.4 is reported alongside it because screening scores one allele at a time and so is the deployment-relevant view, not because the pooled result is unresolvable. **The Spearman row is a rank correlation against** $-\log_{10}$ **IC**$_{50}$, so a *higher* value means better recovery of the continuous affinity ordering and the positive sign is the expected one; computed against raw IC$_{50}$ the same quantity would be negative, which is why the convention is stated rather than left to the reader. It carries no interval, because the evaluation harness reports an analytic p-value for the correlation and not a bootstrap interval (§5.6.2). The "Verdict" column records the outcome of the three-condition contract (§5.6) where that contract applies, and the point estimate where it does not (§5.6.2). **Bolding in the three predictor columns is row-wise**: it marks the highest of the three on that benchmark and metric. Comparing bold marks down a column is not meaningful here, because each row is a different benchmark and metric. **The Δ column is deliberately not bolded**: emphasising the three favourable deltas and not the three unfavourable ones would be a presentational claim rather than a reading of the numbers. 69

6.2 Per-length breakdown on TransPHLA EL (AUPR), from the stored raw predictions of the base five-fold EL head. Deltas are computed from the unrounded scores and may therefore differ by one unit in the last place from the difference of the four-decimal columns shown. Reading the table: DeepNeo-CL leads on 8mers (+0.0075) and 11mers (+0.0100), is level on 10mers, and trails on the dominant 9mer slice (−0.0032). Because 9mers are 71.5 % of the set, the frequency-weighted AUPR delta across these four rows is *negative* (−0.0012): the gains at the non-canonical lengths *offset* the 9mer deficit rather than overturning it. That figure is a summary of this table and **not** a reconstruction of the pooled result: AUPR does not decompose over strata as a frequency-weighted mean, because pooling interleaves positives and negatives from different lengths in one ranking and so changes the curve. The pooled EL AUPR delta is −0.0002 (0.9714 vs 0.9716, §6.2), six times smaller than the weighted figure, which is itself evidence that the two are different quantities. Only the shared sign carries across. **Bolding is row-wise**, marking the better predictor at each length. The pooled EL AUROC advantage (+0.0015, §6.2) is a different metric again and is not decomposed by this table. The benchmark contains **only 8–11mers**, so no row beyond 11 exists and the thesis makes no claim about the 12–14mer range that motivated the length gap in §1.3. 70

6.3 Effect of the label-free BA+EL rank fusion on TransPHLA, with the symmetric comparator. Reading the table: the BA head alone reaches only 0.9546 pooled AUROC — below MHCflurry-2.0 — and the EL head alone 0.9697, so the fused 0.9694 BA figure is a genuine combination effect rather than the BA head's own performance, and the thesis reports it as such. The last rows apply the *same* recipe to NetMHCpan-4.1's two heads as a transparency check, since a scoring recipe that only ever helps the proposed model would be suspect; it is reported whatever it shows. **Bold marks the column maximum, wherever it falls** — which is a NetMHCpan row in all three columns, outright in two and tied in pooled AUROC. Bolding that tracked the proposed model rather than the best value would be a presentational claim the numbers do not support: DeepNeo's 0.9714 AUPR sits against a column maximum of 0.9719, and its 0.9734 per-allele AUROC against a maximum of 0.9750. The bold row *label* is a separate marking and identifies the deployed configuration. 74
6.4 TransPHLA BA fusion table 74
6.5 Effect of the fusion on the IEDB BA static snapshot. Reading the table: the fusion *lowers* pooled AUROC (0.8279 → 0.8144) while *raising* AUPR (0.7365 → 0.7408), so on this benchmark it is not uniformly beneficial and is reported as a mixed result rather than a gain. This divergence is itself informative: a recipe that helps pooled ranking on a 161,652-row benchmark and not on a 1,365-row one is consistent with the cross-allele comparability problem identified in §6.4. 75
6.6 IEDB fusion probe 75
6.7 Mean per-allele AUROC on IEDB BA, grouped by benchmark-set frequency. Reading the table: DeepNeo-CL trails NetMHCpan-4.1 in both subsets, by −0.030 on common alleles and −0.021 on rare ones. The gap is therefore *narrower* on rare alleles, which is the direction the contrastive refinement predicts (RQ2), but both differences sit within roughly one bootstrap standard error at this sample size and neither constitutes a win. The internal ablation evidence for rare-allele benefit is the +0.022 AUPR of §7.4, measured against DeepNeo-CL's own pre-refinement baseline rather than against NetMHCpan-4.1; the two must not be conflated. . . . 77
6.8 Operational positive predictive value in the top 1 % and 5 % of ranked candidates — the regime a vaccine-design pipeline actually operates in, since only the highest-ranked peptides are synthesised. Reading the table: on IEDB the methods are comparable at 1 % while NetMHCpan-4.1 is clearly ahead at 5 % (0.912 vs 0.809), so the practical advantage at shortlist depth belongs to NetMHCpan-4.1 on this benchmark. On TransPHLA both exceed 0.99 and the metric saturates, which is why it is reported as supporting rather than headline evidence. 78

6.9 External head-to-head against post-2020 predictors on identical post-cutoff rows from the official IEDB weekly benchmark, every method scored with its published weights and no retraining. Reading the table: DeepNeo-CL is mid-pack on per-reference AUROC — the official within-allele metric — and marginally ahead on pooled AUROC, and the differences across the leading methods are small relative to the per-pair confidence intervals. These absolute values are *not* comparable with the pooled figures of §6.2, which use a different corpus and protocol; only same-rows, same-protocol values may be compared. Each paired test uses the shared reference–allele subset for that pair. 80
6.10 Post-2020 head-to-head on two-year auto_bench 80
6.11 Statistical-significance summary for every headline comparison, under the three-condition contract of §5.6. The two p-value columns are not interchangeable: DeLong tests pooled cross-allele AUROC, while Wilcoxon tests within-allele consistency, and the contract requires them to agree in direction before a difference is called real. Reading the table: the two TransPHLA rows satisfy both conditions that were evaluated — condition (a) was never computed for any comparison, so no row here satisfies all three (§5.6.2) — and the IEDB pooled row is significant *against* DeepNeo-CL and is reported as such. Note that multiple-comparison correction is not applied across this family — a limitation recorded in §9.12; the TransPHLA tests are far enough from the threshold that correction would not change their verdicts, but that is an argument about these particular tests and not a general defence. 80

7.1 Cumulative ablation ladder on the TransPHLA EL independent benchmark (161,652 rows, 5-fold ensemble). Each row adds one component to the row above it, so the marginal column is that component's contribution *given* the components already present. This is the deployment ladder and it ends at the SupCon v2 base ensemble; the withdrawn transductive pocket rescorer is not a rung (§7.5). 84
7.2 Disposition of every ablation experiment in the Stage 1 log. Seven architectural or training components were retained; twelve experiments were rejected, not promoted, deferred, or withdrawn, and are documented rather than discarded. Ladder figures are marginal steps from Table 7.1. 94

A.1 The four reproducibility layers, in increasing cost. L1 and L2 need no GPU; L3 re-derives the statistics file and L4 retrains from the upstream corpus. Note that L2 confirms checkpoint integrity and verifies no reported metric — re-deriving a number is L3. 125

B.1 The two Stage 2 sub-stages, and which of them this thesis actually delivers. Stage 2B (immunogenicity) is the contribution; Stage 2A (TCR specificity) is scaffold only and carries no result anywhere in this thesis. 135
B.2 Stage 2B training configuration. The decisive row is the first: only the immunogenicity head is trained and the Stage 1 trunk stays frozen, which is what keeps the number of fitted parameters small relative to the supervised stream. 139

B.3 Stage 2B wall-clock budget. The whole stage costs minutes rather than GPU-hours because the trunk is frozen and only a compact head is trained; the dominant cost is scoring the reference predictor, not training. 139
B.4 The three parameter-update strategies considered for Stage 2B, with the trainable footprint of each. The accepted option is the smallest; the rejection of a full fine-tune is an over-fitting judgement, not a measured result. Note that the n in the third column is the TESLA set, and the supervised training stream is BigMHC `im_train.csv` (Table B.5) — the column is a surviving annotation from the TESLA-era design and is not the size the head was fitted on. 140
B.5 Stage 2B benchmark splits, all from BigMHC's published release. Only `im_test` carries a headline number; `im_val` is used for checkpoint selection and is therefore not an independent test. 140
B.6 Metrics reported for Stage 2B and the uncertainty attached to each. The asymmetry is deliberate and is the same one §5.3 records: the two discriminative metrics carry bootstrap intervals and the operational metrics are point estimates. 141
B.7 Published TCR-specificity predictors a Stage 2A effort would have to compare against. The right-hand column carries no comparable figure for any of them, which is a property of the published literature rather than an omission here: none reports a result on a task matching the one scoped above. It is not the reason Stage 2A is deferred — that reason is the data-volume argument given earlier in this section — and a Stage 2A benchmark would in any case be run on the IEDB T-cell-assay union rather than on TESLA. 143
B.8 Primary Stage 2B head-to-head on BigMHC `im_test`. **Bold marks the column maximum, which is a BigMHC variant on both metrics**; the bold row label identifies this thesis's model. Bolding DeepNeo's 0.5735 AUROC against a column maximum of 0.5974, or its 0.3099 AUPR against 0.3248, would read as a claim the numbers do not support. Every pairwise comparison here is non-significant after Holm correction, so the honest reading of the table is parity rather than a ranking. 143

C.1 Controlled documentation set referenced by the thesis. 152

D.1 Two-year pooled IEDB auto_bench standing 155
D.2 Head-to-head ranking on the two-year auto_bench corpus 159
D.3 Paired Wilcoxon: BA+EL fusion vs each post-2020 competitor 160

List of Equations

3.1 Decontaminated IEDB benchmark set . . . 21
3.2 Scaled dot-product attention . . . 30
3.3 Masked-language-model pre-training objective . . . 30
3.4 Peptide-conditioned cross-attention over the pseudo-sequence . . . 31
3.5 Censored mean-squared error for binding affinity . . . 31
3.6 Focal binary cross-entropy for presentation . . . 32
3.7 Joint Stage 1 binding-and-presentation objective . . . 32
3.8 Within-allele pairwise ranking auxiliary loss . . . 32
3.9 Label-distribution-smoothing sample weight . . . 32
3.10 Effective-number subgroup weight . . . 33
3.11 Stage 1 per-sample objectives, one per task batch . . . 33
3.12 Supervised contrastive loss . . . 34
3.13 Stage 1 objective with contrastive refinement . . . 35
3.14 The 370-dimensional pocket feature vector (withdrawn experiment) . . . 36
3.15 Pocket-rescored affinity (withdrawn experiment) . . . 36
4.1 Asymptotic cost of one forward pass, by component . . . 47
5.1 Area under the ROC curve, with the tie convention . . . 55
5.2 Area under the precision–recall curve . . . 55
5.3 Spearman rank correlation . . . 56
5.4 Positive predictive value at rank k . . . 56
5.5 DeLong structural component V_{10}, over the positives . . . 57
5.6 DeLong structural component V_{01}, over the negatives . . . 57
5.7 DeLong variance of the AUROC difference . . . 57
5.8 Five-fold ensemble score . . . 62
B.1 Stage 2B input: the frozen Stage 1 representation . . . 137
B.2 Immunogenicity head . . . 138
B.3 Focal loss for immunogenicity . . . 138
B.4 Stage 2B head-to-head difference and its DeLong p . . . 143

List of Abbreviations

Abbreviation	Definition
AUPR	Area under the Precision–Recall curve
AUROC	Area under the Receiver-Operating-Characteristic curve
BA	Binding affinity
BCE	Binary cross-entropy
$CD8^+$	Cytotoxic T-cell
CI	Confidence interval
CPT	Continued pre-training
CV	Cross-validation
DeepNeo-CL	The model developed in this thesis: a contrastively refined, pan-allele peptide–MHC class I predictor
DeLong	DeLong's paired AUROC significance test
EL	Eluted ligand
EMA	Exponential moving average (of model weights)
ESM2	Evolutionary Scale Modeling, version 2 [2]
FDR	False discovery rate
GPU	Graphics processing unit
HLA	Human leucocyte antigen
HLA-I	HLA class I (HLA-A, -B, -C)
IC50	Half-maximal inhibitory concentration
IEDB	Immune Epitope Database [3]
LDS	Label-distribution smoothing [4]
LM	Language model
LoRA	Low-rank adaptation
MD5	Message-Digest Algorithm 5 (file hash)
MHC	Major histocompatibility complex
MLM	Masked language modelling
MLP	Multi-layer perceptron
MS	Mass spectrometry
NaN	Not-a-number (numerical pathology)
OHEM	Online hard example mining [5]
pLDDT	predicted Local Distance Difference Test (AlphaFold confidence)
pMHC	Peptide–MHC complex
pMHC-I	Peptide–MHC class I complex
PPV	Positive predictive value
PSSM	Position-specific scoring matrix
PTM	Post-translational modification
ROC	Receiver-Operating-Characteristic curve
RNG	Random-number generator
SRCC	Spearman rank correlation coefficient
SHA-256	Secure Hash Algorithm 256-bit
SupCon	Supervised contrastive loss [6]
TAP	Transporter Associated with Antigen Processing
TCR	T-cell receptor
TOR	TransPHLA / Outer-benchmark
TPR/FPR	True / false positive rate
UniRef50	UniProt Reference Clusters at 50% sequence identity
V37 / VVSFZL37	37-dimensional position-resolved physicochemical residue descriptor
VRAM	Video random-access memory (GPU memory)
xattn	Cross-attention

Notation

Symbol	Meaning
$p = (p_1, \ldots, p_L)$	Peptide string of length $L \in \{8, \ldots, 14\}$
$h = (h_1, \ldots, h_{34})$	HLA pseudo-sequence (34 residues)
$\mathbf{X}^{(p)} \in \mathbb{R}^{L \times d}$	ESM2-35M peptide token embeddings, $d = 480$
$\mathbf{X}^{(h)} \in \mathbb{R}^{34 \times d}$	ESM2-35M pseudo-sequence token embeddings
$\mathbf{H}^{(p \leftarrow h)}$	Cross-attention output (peptide queries over pseudo keys/values)
$\psi_{370}(p,h) \in \mathbb{R}^{370}$	Crystal-template pocket descriptor
$s_{\text{base}}(p,h)$	Base 5-fold ensemble logit
$s_{\text{final}}(p,h)$	Rescored logit (withdrawn experiment; not in deployment headline)
$\hat{y}, y$	Predicted and ground-truth IC50 target
$\hat{p}, p_t$	Predicted probability, true-class probability
$\sigma(\cdot)$	Logistic sigmoid function
τ	SupCon temperature hyperparameter ($\tau = 0.2$)
γ	Focal-loss focusing parameter ($\gamma = 2$)
β	Cui 2019 effective-number β ($\beta = 0.999$)
α	Reranker base-logit weight (learnable, initialised at 1.0)
L	Peptide length
B	Number of bootstrap resamples
n, n_+, n_-	Benchmark size; positive count; negative count
$\sigma(\text{AUPR})$	Bootstrap standard deviation on AUPR
T_{net}	Union of NetMHCpan-4.1 5-fold training `(peptide, hla)` pairs
E'	Decontaminated IEDB benchmark set
g	Joint subgroup index for effective-number weighting
$\mathscr{L}_{\text{MLM}}$	Masked-language-modelling loss
$\mathscr{L}_{\text{BA}}$	Censored mean-squared-error binding-affinity loss
$\mathscr{L}_{\text{EL}}$	Focal binary cross-entropy presentation loss
$\mathscr{L}_{\text{rank}}$	Within-HLA pairwise ranking hinge loss
$\mathscr{L}_{\text{SupCon}}$	Khosla 2020 supervised contrastive loss
$\mathscr{L}_{\text{S1B}}, \mathscr{L}_{\text{S2}}$	Stage 1B and Stage 2 total objectives
$\lambda_{\text{BA}}, \lambda_{\text{EL}}, \lambda_{\text{rank}}, \lambda_{\text{SupCon}}$	Loss weights (0.6, 0.4, 0.4, 0.1 respectively)

A glossary of natural-language terms is given in the project methodology documentation (Appendix C).

Author Statement of Responsibility

The author takes full scientific, methodological, and editorial responsibility for this thesis. All reported results, numerical values, and statistical claims are traceable to the documented experiment outputs and cited references.

Chapter 1

Introduction

Stage naming. This document uses the canonical taxonomy defined in `STAGE_NAMING.md` v2 (2026-05-21). Stage 1 = pMHC binding/presentation (1A = BA, 1B = EL, plus SupCon refinement); **Stage 2B = peptide–MHC immunogenicity** (Appendix B of this thesis); Stage 2A = TCR-specificity (future work).

1.1 Motivation

Each year more than 18 million people are newly diagnosed with cancer worldwide [7]. Of the therapeutic strategies that have moved from bench to bedside in the last decade, *personalised cancer immunotherapy* — vaccines and adoptive T-cell therapies that target the precise mutations carried by a single patient's tumour — has shown the most consistent translational signal [8, 9]. The molecular substrate of every such therapy is the **peptide–major-histocompatibility-complex class I complex (pMHC-I)**: 8- to 14-residue peptide fragments derived from intracellular proteins, loaded onto the cleft of a major histocompatibility complex (MHC) class I molecule — in humans, a human leucocyte antigen (HLA) of the A, B or C locus — and displayed at the cell surface for surveillance by $CD8^+$ T-cells.

The biology of T-cell-mediated tumour recognition imposes three sequential filters on a candidate peptide. A peptide that arises uniquely from a tumour mutation is a *neoantigen* and a candidate vaccine epitope only if it (1) **binds** the patient's HLA with sufficient affinity (BA), (2) is **processed and presented** on the cell surface (EL), and (3) the resulting pMHC complex is **recognised** by a $CD8^+$ T-cell receptor (IM). These three filters are not independent prediction tasks; they share substrate (the same peptide and HLA), share biophysical constraints (binding-pocket geometry and anchor residues), and are partially predictive of one another. The central thesis-architecture decision of DeepNeo-CL is therefore to model the pathway end-to-end with one shared representation: **Stage 1** trains the BA + EL halves jointly on a shared trunk (Chapters 3–9), and **Stage 2B** extends the same trunk with a compact immunogenicity head (Appendix B) so that a single forward pass returns BA, EL, and IM scores for a (peptide, HLA) input.

1.1.1 Stage taxonomy at a glance

Throughout the thesis, references to Stage 1, Stage 2B, and Stage 2A follow Table 1.1, which is the single place the taxonomy is defined.

Whole-exome sequencing of a single tumour typically generates 10^2 to 10^4 candidate mutant

Stage	Task scope	Where in this thesis
Stage 1A	BA (binding affinity)	Chapters 3–9
Stage 1B	EL (presentation)	Chapters 3–9
Stage 2B	IM (peptide–MHC immunogenicity)	Appendix B
Stage 2A	TCR specificity (future work)	n/a (scaffold only)

Table 1.1: Canonical stage taxonomy used throughout the thesis. The right-hand column is resolved by cross-reference rather than written out, so it cannot fall out of step with the chapter numbering again.

peptides per HLA allele. Wet-lab validation of even a small fraction is prohibitive; therefore the entire vaccine pipeline is rate-limited by the **per-pair false-positive rate of the upstream binding predictor**. Improving prediction quality by even a small absolute margin compounds across thousands of candidate peptides into a sharply reduced experimental burden — and a faster, cheaper personalised vaccine.

Why further work is needed when a good predictor already exists. The obvious objection to this thesis is that the problem looks solved: the reference predictor scores close to 0.96 on the area under the receiver-operating-characteristic curve (AUROC) on standard benchmarks (§1.2), and a remaining margin of a few thousandths is not obviously worth a thesis. Two things are wrong with that reading, and together they are the reason for this work.

The first is that an *aggregate* score is the wrong summary for this application. A patient does not present with the average allele; they present with their own, and a predictor that is excellent on HLA-A*02:01 — which dominates both the training corpora and the benchmarks — can be materially worse on an allele for which only a few hundred measurements exist. Aggregate accuracy conceals exactly the regime in which a clinical pipeline is most likely to fail, and the patients it fails are those with less-studied haplotypes. That is a distributional problem, not a capacity one, and it is what §1.3's first gap names.

The second is that the standard of evidence in this literature does not currently support fine comparisons. Margins of a few thousandths are being claimed in a field where confidence intervals are often absent, single folds are sometimes used as headlines, and decontamination against the dominant training corpus is not always performed. Improving a predictor is of little use if the improvement cannot be told apart from noise, so a second need is methodological: a comparison protocol strict enough that a small margin means something, applied even where it costs the thesis its own headline. Both needs are addressed here, and the second is the one this work defends most confidently.

1.2 The state of pMHC-I prediction in 2026

For nearly two decades the reference predictor has been **NetMHCpan** [10, 11]. NetMHCpan-4.1 trains a shallow allele-aware feed-forward network on the joint binding-affinity (BA) and mass-spectrometry eluted-ligand (EL) corpora, scores at an AUROC of ≈ 0.96 on standard external benchmarks, and remains the single most widely cited baseline in the neoantigen-discovery literature. **MHCflurry-2.0** [12] is a stronger BA-only predictor in some independent evaluations and weaker in others. Several transformer-based predictors have appeared since 2022 — including **TransPHLA** [13], **MHCfovea** [14], **BigMHC** [15], and **ImmuneApp** [16] — but none, in our

Chapter 2 survey, were shown to surpass NetMHCpan-4.1 on a strictly external benchmark under the full statistical protocol used here (paired testing with confidence intervals and decontamination controls). Appendix D §D.5 verifies this qualitative reading on independent post-cutoff data by scoring all four on the same 1,456-row IEDB auto_bench corpus: on the deployment-fair within-allele metric, DeepNeo-CL's BA+EL fusion is statistically indistinguishable from every one of the four and from NetMHCpan-4.1.

An informal survey of recent transformer-based pMHC-I predictor papers (full discussion in Chapter 2) indicates that confidence intervals on the headline metric are rarely reported, that single-fold numbers are frequently used as the headline, and that pair-level decontamination against the dominant training corpus (NetMHCpan-4.1) is not always performed. The survey behind these observations is *structured* rather than systematic, and the distinction is worth stating precisely because the claims above are field-wide ones. Its inclusion and exclusion criteria are given at §2.1: class I predictors evaluated on at least one external benchmark, published or preprinted between 1995 and 2025. What it does *not* have is a reproducible search protocol, a screened candidate set with counts at each stage, or two independent reviewers — so the observations are best read as characterising the predictors this thesis actually compares itself against, and should be treated cautiously as claims about the field as a whole. Quantifying them properly would be a systematic review, which is a piece of work in its own right and is not attempted here. Nevertheless, the methodological posture of this thesis is a deliberate response to this pattern: every claim is reported with a bootstrap-95% confidence interval, every headline number is a 5-fold ensemble, and every benchmark is pair-level decontaminated against the NetMHCpan-4.1 training corpus.

1.3 Three gaps in current pan-allele predictors

We identify three concrete gaps in the NetMHCpan-4.1 architecture that motivate a deeper model:

1. **Long-tail HLA allele transfer.** The official NetMHCpan-4.1 5-fold corpus contains a sizeable tail of low-frequency alleles (< 100 training pairs each). Shallow allele embeddings under-fit this tail; a regularised representation should transfer.

2. **Length extremes.** Peptides of length 10–14 follow systematically different anchor rules to canonical 9mers, and a single global model under-fits that distribution. (Throughout this thesis an *nmer* denotes a peptide of exactly *n* amino-acid residues. The *canonical* MHC class I peptide length is 9; the closed groove imposes a practical upper bound near 14 residues, and lengths 10–14 bind with systematically different anchor chemistry, as Chapter 2 §2.8 develops. Note that the independent benchmark used in Chapter 6 covers 8–11mers only, so benchmark coverage and biological range are not the same thing and are kept distinct throughout.)

3. **Structural priors at inference.** The 34-residue MHC *pseudo-sequence* — the structurally-derived map of which $\alpha 1/\alpha 2$ positions contact the peptide — is a known prior on cross-allele transfer, and every pan-allele predictor consumes it. What none of them does is make it *position-addressable*: NetMHCpan-4.1 encodes the 34 residues and concatenates them into a shallow network, so the model can condition on the groove as a whole but cannot attend to the individual residue that constrains a given peptide position. The gap is in the interaction operator, not in the input.

1.4 Research questions

The three gaps of §1.3 are statements about what existing predictors do not do. Turning them into research questions requires asking, for each gap, whether the mechanism proposed to close it actually closes it *and* whether the resulting predictor survives comparison with the established reference under a protocol strict enough for the answer to mean something. Each gap therefore yields one question about a mechanism, and the methodological pattern described in §1.2 yields a fourth question about measurement.

RQ1 — Structural priors. Every pan-allele predictor since 2007 consumes a sequence of groove residues rather than an allele identifier, so the open question is not *whether* to use the pseudo-sequence but how to combine it with the peptide. Does consuming it as a *position-addressable cross-attention target*, encoded by the same protein-language model as the peptide, improve peptide–MHC-I binding and presentation prediction over the flat concatenated embedding the lineage uses? *From gap 3.* Answered by the architecture of Chapter 4 §4.2 and the cross-attention ablation of Chapter 7 §7.1 and §7.3.

RQ2 — Long-tail allele transfer. Can a contrastive regularisation of the representation, grouped by structurally meaningful pseudo-sequence clusters rather than by allele identity, improve transfer to low-frequency HLA alleles without degrading performance on common ones? *From gap 1.* Answered by the SupCon v2 method of Chapter 3 §3.19 and the rare-allele analysis of Chapter 7 §7.4, with the per-allele breakdown in Chapter 6 §6.6.

RQ3 — Length extremes. Does a single pan-length model with position-resolved physicochemical features handle non-canonical peptide lengths, or does the length distribution require separate treatment? *From gap 2.* Answered by the per-length analysis of Chapter 6 §6.3 and the imbalance-correction ablations of Chapter 7 §7.2. Note that the deployed model carries no *length-aware* imbalance correction. The corrections it does carry are label-distribution smoothing on the continuous half-maximal-inhibitory-concentration (IC50) target (E11) and effective-number re-weighting over joint pseudo-cluster × affinity-bin groups (E12–E13), neither of which is keyed on length (Chapter 3 §3.18). Both length-conditioned variants that were tested — E10's length × BA-bin grouping and E9's weight gated on lengths 8–11 — were **rejected for deployment**, the first as a null result and the second as a slice-local diagnostic (§7.2). Those rejections are part of RQ3's answer rather than a premise of its question.

RQ4 — Measurement. Under a declared evidential contract — pair-level decontamination, five-fold ensembling, bootstrap confidence intervals, paired DeLong testing and per-allele Wilcoxon agreement — does the resulting predictor outperform NetMHCpan-4.1 on independent post-cutoff benchmarks, and on which metrics does it not? *From the methodological pattern in §1.2.* Answered by the contract of Chapter 5 §5.6 and the two-sided result of Chapter 6, with the cross-predictor extension in Appendix D.

RQ4 admits a negative answer in part, and it receives one: Chapter 6 reports that DeepNeoCL exceeds NetMHCpan-4.1 on three of four TransPHLA metrics and ties on the fourth, is *not distinguished* from it on within-allele IEDB ranking over two years of auto_bench, and *trails* it on the static IEDB snapshot on pooled and within-allele ranking alike. A question that could only be

answered favourably would not be a research question, and the reader should weigh that the four questions were written down in their present form *after* these results were frozen (§5.6.1 gives the chronology for the evidential standard, and the same applies here): what the unfavourable answer demonstrates is that the thesis reports it, not that it was anticipated. Chapter 10 §10.1 returns to each of RQ1–RQ4 and states what the evidence does and does not establish.

1.5 Thesis statement and scope

We defend the following single defensible claim:

> **DeepNeo-CL**—an ESM2-35M pMHC-I predictor with peptide×HLA pseudo-sequence cross-attention and SupCon v2 representation refinement (Stage 1 pipeline; `docs/model/STAGE_NAMING.md` v2), evaluated as a five-fold ensemble without eval-set fitting—matches or exceeds NetMHCpan-4.1 on the external TransPHLA benchmark: significantly higher BA AUROC/AUPR after label-free BA+EL fusion (0.9694/0.9705 vs 0.9638/0.9642; DeLong $z = 22.9$), higher EL AUROC (0.9697 vs 0.9682; $p = 1.8 \times 10^{-8}$), and tied EL AUPR. On IEDB BA the two predictors are **not distinguished** on within-allele (per-reference) ranking over two years of IEDB auto_bench (Wilcoxon $p = 0.65$) — a failure to detect a difference, not a demonstration of equivalence. On the smaller static snapshot ($n = 1{,}365$) DeepNeo-CL **trails on both measures**: pooled cross-allele aggregation (ΔAUROC -0.043, DeLong $p = 3.2 \times 10^{-8}$) and, by roughly half that margin, within allele (-0.023). Appendix D extends the comparison to BigMHC, MHCfovea, TransPHLA, and ImmuneApp on identical post-cutoff rows.

The thesis does **not** claim:

- Stage 2B headline claims are restricted to BigMHC `im_test`; no broad clinical-immunogenicity claim is made.
- Class II MHC.
- Post-translational modifications.
- TCR specificity.
- Strict dominance over MHCflurry-2.0 on eluted-ligand presentation (no comparable EL API in our pipeline).

1.6 Contributions

This thesis makes three substantive contributions, each independently validated by ablation. The research question each addresses is given in brackets, and Chapter 10 §10.1 reports how far the evidence carries it.

Each contribution is **specified, built, and only then measured**, and the three happen in different places. The specification is in the methodology chapter (Chapter 3) and the implementation in the architecture chapter (Chapter 4) — so a reader meets every contribution within the first third of

the thesis. What necessarily comes later is the *attribution*: the cumulative ablation of Chapter 7 is what assigns a number to each component, and no honest version of that can precede the results it decomposes. The entries below therefore give all three locations, so the deferral is visible rather than discovered.

- **C1 — Architecture.** A peptide-conditioned cross-attention layer over the 34-residue MHC pseudo-sequence inside an ESM2-35M transformer backbone. In the cumulative ablation (Chapter 7, Figure 7.1) the cross-attention step contributes $\Delta\text{AUROC}_{\text{TransPHLA}-\text{EL}} = +0.002$ in isolation, and the pseudo-sequence *family* — concatenation plus cross-attention — contributes $+0.007$ measured after the V37 physicochemical rung. The $+0.012$ span from the peptide-only ESM2 baseline is *not* the family's contribution: it also contains V37's own $+0.005$ (§7.1). The largest single rungs on the ladder are $+0.005$. Specified in Chapter 3 §3.16 and Chapter 4 §4.2; ablated in Chapter 7 §7.3. *[RQ1]*

- **C2 — Training method.** SupCon v2, a supervised contrastive auxiliary loss on the *pseudo-cluster* label (12 clusters), wrapped in fp32 with a numerical NaN guard. Delivered $+0.022$ AUPR on the IEDB rare-allele subset ($n_{\text{train}} < 100$) while not regressing on common alleles—exactly the long-tail regularisation signature predicted by Khosla et al. [6]. It is within noise on pooled TransPHLA EL, so this is a subgroup contribution. Specified in Chapter 3 §3.19; ablated in Chapter 7 §7.4, with the per-allele breakdown in Chapter 6 §6.6. *[RQ2]*

- **C3 — Deployment-safe BA+EL fusion.** A label-free rank-average of the two frozen heads raises pooled TransPHLA BA to 0.9694 (vs NetMHCpan-4.1 0.9638; no eval-set fitting), providing a like-for-like comparator to NetMHCpan's joint BA+EL design. An exploratory crystal-template pocket rescorer (Chapter 7 §7.5) showed transductive in-sample gains only and is excluded from all deployment claims. Specified in Chapter 3 §3.21; evaluated in Chapter 6 §6.5. *[RQ4]*

A fourth, methodological contribution is the **milestone reproducibility vault** released alongside this thesis: an immutable, air-gapped artefact containing the source code, 5-fold ensemble weights with Message-Digest Algorithm 5 (MD5) integrity hashes, the raw true-label and predicted-score arrays, 12 publication-grade figures, and a single-pair forward-pass validator that reproduces the headline numbers in under two minutes of CPU time. At submission time, this is among the most complete publicly shipped reproducibility artefacts for a pMHC-I predictor thesis workflow.

1.7 Roadmap

The remainder of the thesis is organised as follows, with the research question each part addresses given in brackets.

- **Chapter 2** surveys prior pMHC-I predictors — from position-specific scoring matrices through the NetMHCpan pan-allele paradigm, MHCflurry-2.0, transformer approaches, and protein foundation models — and identifies the methodological gap this thesis responds to. *[RQ4]*

- **Chapter 2** develops the biological background: the MHC-I antigen presentation pathway, the canonical anchor model, the structural meaning of the 34-residue pseudo-sequence, and the long-tailed allele frequency distribution. *[RQ1, RQ2]*
- **Chapter 3** specifies the data: the NetMHCpan-4.1 official five-fold training corpus and its stratification invariants (§3.4); the three independent benchmarks — decontaminated IEDB BA, TransPHLA, and TESLA for zero-shot immunogenicity — with the pair-level decontamination protocol and a statistical-power analysis for each (§3.5–4.5); the design of a homology-reduction sensitivity analysis, which was specified and not performed (§3.9); and the provenance record (§3.11). *[RQ4]*
- **Chapter 3** develops the methodology and the mathematics of the model: input representation, the physicochemical projection, cross-attention, the multi-task objectives, imbalance correction, contrastive refinement, and the inference pipeline. *[RQ1, RQ2, RQ3]*
- **Chapter 4** turns the methodology into an implementation: a component-level inventory with parameter counts, hyperparameter justifications traced to ablations, an architectural comparison against the reference predictors, and the complexity, memory and scalability analysis (§4.6). *[RQ1]*
- **Chapter 5** defines the evaluation metrics and the three-condition statistical contract, together with its provenance and the respects in which the thesis did not meet it (§5.6.1, §5.6.2). *[RQ4]*
- **Chapter 6** presents the results with bootstrap confidence intervals, paired significance tests, and per-allele and per-length breakdowns. *[RQ1–RQ4]*
- **Chapter 7** reports the cumulative ablation ladder and the twelve experiments that did not earn a place in the deployed model. *[RQ1, RQ2, RQ3]*
- **Chapter 8** interprets where DeepNeo-CL separates from NetMHCpan-4.1 and where it does not. *[RQ1–RQ4]*
- **Chapter 9** states the limitations, and **Chapter 10** scopes future work.
- **Chapter 10** revisits each research question and states what the evidence establishes.
- **Appendix A** is the reproducibility appendix, documenting the milestone vault.
- **Appendix B** reports the Stage 2B immunogenicity extension, which is outside the Stage 1 claim of §1.5 and is reported as parity-with-caveats against BigMHC.
- **Appendix D** extends the comparison to BigMHC, MHCfovea, TransPHLA and ImmuneApp on identical post-cutoff rows from the official IEDB weekly benchmark.

Chapter 2

Literature review and background

2.1 Scope of the review

This chapter surveys the methodological lineage that the present thesis builds on and the methodological gap it fills.

Inclusion criteria. A predictor is in scope if it (i) predicts peptide–MHC *class I* binding affinity or presentation, (ii) has been evaluated on at least one benchmark external to its own training corpus, and (iii) was published or released as a preprint between 1995 and 2025. The second criterion is the one that does most of the work: a predictor evaluated only on a held-out split of its own corpus cannot be compared with one evaluated externally, and admitting both would make the lineage below a comparison of evaluation protocols rather than of methods.

Exclusions, and why each is excluded. The boundaries are drawn where a comparison would stop being informative rather than where the literature stops being interesting:

- **Class II MHC predictors.** The class II groove is open at both ends, so peptides are variable-length and the anchor geometry that class I prediction exploits does not apply. A method that works well there tells us little about class I, and the two sub-fields have largely separate benchmark communities, so cross-citation of results would be misleading rather than merely incomplete.
- **Immunogenicity and T-cell-recognition predictors that emit no binding score.** These answer a later question in the same pathway — will a presented peptide be *recognised* — and are therefore not comparable on the binding and presentation metrics this thesis reports. They are excluded from the *review*, not from the thesis: the same pathway motivates the Stage 2B work in Appendix B.
- **Docking and structure-prediction tools not paired with a binding-label benchmark.** These produce a structure or a score with no labelled binding outcome to validate against, so there is no metric on which to place them beside the predictors surveyed here. The exclusion is about the absence of a shared yardstick, not about structural methods being irrelevant — one such method was tried in this thesis and withdrawn (§7.5).
- **Cohort-level retrospective vaccine-pipeline audits.** These evaluate an entire clinical pipeline, of which a binding predictor is one component among many, so a result cannot be attributed to the predictor. They bear on whether this work matters and not on how it compares.

2.2 Evolution of the field

2.2.1 Position-specific scoring matrices (1995–2007)

The earliest serious pMHC-I predictors were per-allele **position-specific scoring matrices (PSSMs)**: BIMAS [17], SYFPEITHI [18], and several follow-on matrices. Each was hand-curated from at most a few hundred binders per allele, and could not generalise across alleles. The field hit a hard ceiling at this level because every new allele required new training data.

2.2.2 Pan-allele neural era (2007–2020)

NetMHCpan [10, 11, 19–21] introduced the **pan-allele abstraction**: input the 34-residue MHC *pseudo-sequence* alongside the peptide and let a shallow feed-forward network learn the cross-allele transfer rule. The pseudo-sequence is the canonical structurally-derived map of the $\alpha 1/\alpha 2$ residues that both lie within 4 Å of a bound peptide in the aligned crystal structures *and* vary across alleles — the polymorphism criterion that reduces the contact set to 34 positions (§2.13) — a *structural* prior, not a sequence-similarity prior. NetMHCpan-4.1 [11] is the current reference implementation, trained jointly on BA + EL data with allele-aware allele embeddings.

2.2.3 MHCflurry-2.0

MHCflurry-2.0 [12] is a deeper neural-network predictor trained on a curated BA + EL union, with hand-curated allele embeddings and an attention-free architecture. It often surpasses NetMHCpan on EL slices and ties or loses on BA — but the EL gain is contingent on the specific mass spectrometry (MS) dataset chosen for evaluation.

2.2.4 Transformer era (2022–)

The 2022 release of TransPHLA [13] introduced self-attention over the peptide sequence, and kept the pan-allele convention of describing the allele by its 34-residue pseudo-sequence — the same table this thesis reads — encoded through its own residue embedding rather than through a protein language model. It shipped a 161,652-row pan-allele independent benchmark that we adopt in Chapter 5, and reports BA AUROC 0.95+ on it. MHCfovea [14], a 2021 convolutional-network ensemble rather than a transformer, is noted here for architectural contrast: it reads a longer 182-residue binding-cleft sequence, and refined accuracy and interpretability, but did not release a reproducibility artefact. Several ESM-based predictors (e.g. ESM-MHC pre-prints) have appeared but generally report only single-fold numbers without bootstrap CIs.

2.2.5 Foundation-model era

Recent work explores **protein language models** as a pretraining substrate: ESM2 [2] and its descendants provide evolutionarily-conditioned embeddings that transfer to many downstream tasks. The ESM2-35M backbone we use here was selected to fit the 24 GB of video random-access memory (VRAM) on an RTX-4090 (Phase 1 redux on ESM2-150M is scoped but deferred — see Chapter 10).

2.2.6 The lineage assessed: mechanism, contribution, limitation, suitability

The subsections above describe what each generation did. This one assesses them, because a review that only narrates is not yet a review: for each, what problem it solved, what it cannot do, and whether this thesis can use it. Table 2.1 is the summary; the paragraphs after it take up the three judgements that bear directly on the design adopted here.

The pseudo-sequence is the device the lineage converged on, and what distinguishes the generations is how the two sides are combined. The pan-allele networks' central idea was to describe an allele by the groove residues that touch the peptide, so that a model can score an allele it has never trained on, and §2.3.1 records it as the single most consistently rewarded device in the lineage. It is worth being exact about how durable that convergence is, because it bounds what any later architecture can claim: NetMHCpan takes the 34-residue pseudo-sequence, MHCflurry-2.0 maps an allele to an amino-acid sequence and encodes those residues, MHCfovea takes a longer 182-residue cleft sequence, and TransPHLA takes the *same* 34-residue table this thesis reads. No generation after 2007 describes an allele by an identifier, so no later design can be motivated as restoring a structural description that was lost.

What is left to differ is the interaction, and that is where this thesis sits. Two things separate the architecture of Chapter 4 from the comparators, and both are narrower than a new input. First, the peptide and the groove pass through the *same* pretrained protein-language encoder, so the two representations are commensurable and the groove side inherits structure learned from general protein sequence; every comparator embeds the two sides with separate modules trained on pMHC pairs alone. Second, the interaction operator is peptide-conditioned cross-attention against a position-addressable groove rather than a concatenation, which is what allows a gradient to reach the individual groove residue responsible for a preference (§2.3.6). The cumulative ablation of Chapter 7 prices the interaction operator — $+0.002$ AUROC for cross-attention in isolation and $+0.007$ for the pseudo-sequence family measured after the physicochemical rung — while the first difference, sharing one pretrained encoder across both sides, is a design choice argued in Chapter 4 rather than a rung with a measured delta of its own. That measured figure for the interaction, not a claim of restoring a discarded prior, is the contribution this thesis makes at the input.

The PSSM generation is unsuitable as a baseline but remains useful as evidence. Its scoring model — independent positions — is now known to be wrong, and no comparison against it would be informative. What survives is the empirical content: the allele-specific anchor motifs it catalogued are still the clearest statement of *which* positions matter, and this thesis relies on that in two places: the anchor-gated branch of §4.2, which learns a gate over an explicit anchor-position indicator, and the interpretation of the per-length result in Chapter 8. The physicochemical descriptor itself is computed position by position across the whole peptide rather than at the anchors only; the one component that encoded P2 and PΩ specifically was the crystal-template pocket rescorer, which is withdrawn.

Two limitations in the lineage are methodological rather than architectural, and they are the ones this thesis responds to. Reported comparisons are frequently made against figures measured on a different corpus, and reproducibility artefacts are rarely released in a form that permits an independent check. Neither is a property of any method above; both are properties of how the field reports. §2.4 takes this up, and it is the gap this thesis claims most confidently to fill.

Table 2.1: The pMHC-I prediction lineage assessed. "Usable here" is the operative column: a method is usable if this thesis can either build on its mechanism or compare against it on identical rows. Each cell records the primary use; the pan-allele shallow nets are the one generation the thesis does both with — NetMHCpan-4.1 supplies the pseudo-sequence device (§2.3.1) and is the reference comparator — while the PSSMs supply motif evidence whose scoring model is superseded, and the language models supply a trunk that cannot itself be compared against. Note that from the pan-allele generation onward every row that describes an allele does so by a sequence of groove residues rather than an identifier; the protein-language-model row describes no allele at all, its trunk being allele-agnostic.

Generation	Mechanism	Problem it solved	Principal limitation	Usable here
PSSMs [17, 18]	A per-allele position-weight matrix scoring each residue independently	Made the anchor motif explicit and computable	Positions treated as independent, and a new allele needs new data, so no cross-allele transfer	As motif evidence only
Pan-allele shallow nets [10, 11]	A feed-forward network taking the peptide *and* the 34-residue pseudo-sequence	Cross-allele transfer: one model for alleles with no training data of their own	Limited capacity to model long-range or higher-order interactions within the peptide	Yes — reference comparator
Attention-free deep nets [12, 14]	Deep MLP (MHCflurry-2.0) or CNN ensemble (MHCfovea) over the peptide and a per-residue encoding of the allele's groove sequence	Showed that added depth and better data curation still pay, and that the groove encoding can be read for interpretability	No mechanism for long-range or higher-order interaction within the peptide, and the two sides are combined by a fixed encoding rather than a learned interaction	Partly — BA only
Transformer predictors [13]	Self-attention over the peptide; the allele still supplied as its 34-residue pseudo-sequence, through a separate task-trained embedding	Modelled within-peptide position interactions directly	Peptide and groove are embedded by separate modules trained on pMHC data alone, so neither side carries representation learned from general protein sequence	Yes — benchmark and comparator
Protein language models [2]	Masked-language pretraining over a large protein corpus	A general sequence representation transferable without task labels	The objective never sees peptide–MHC *pairs*, so pair-specific structure must be added downstream	Yes — adopted as the trunk

2.3 Architectural devices that have worked

2.3.1 Pseudo-sequence concatenation

The single most important architectural choice across all pan-allele predictors is to consume the 34-residue MHC pseudo-sequence rather than a one-hot allele identifier. This is the *structurally meaningful* prior on cross-allele transfer [10], and it is what allows a model to score an allele for which it has no training data at all: two alleles with nearly identical grooves are near-identical inputs, where to an identifier they would be as unrelated as two arbitrary labels. The device has been retained by every generation since, which is the strongest available evidence for it; this thesis runs no one-hot allele ablation of its own, so no figure is quoted here for the size of the effect.

2.3.2 Multi-task BA + EL joint training

NetMHCpan-4.1 and MHCflurry-2.0 both train a shared trunk with a BA regression head and an EL classification head. Empirically the joint signal regularises both heads — EL data has wider coverage, BA data has more precise labels.

2.3.3 Position-resolved physicochemical features

VVSFZL37 (the 37-dimensional VVSFZL family of residue-resolved physicochemical descriptors) and its variants have been shown to contribute a $+0.02$–$+0.03$ AUROC lift over a pure language-model representation (in our work, ablation E3, Chapter 7). The intuition is that anchor residue chemistry (volume, hydrophobicity, flexibility) is the dominant biochemical determinant of binding, and a pure LM embedding does not isolate this signal.

2.3.4 Imbalance correction

Effective-number weighting [22] and label-distribution smoothing [4] are the two regression-side imbalance correctors with documented positive results in our ablations (E11, E12, E13 — see Chapter 7). Sampler-side correction (OHEM-style allele hard-mining, focal binary cross-entropy on the EL task) was less consistently positive (E5–E7 — rejected).

2.3.5 Contrastive regularisation

Supervised contrastive learning [6] is a natural regulariser for an embedding space with a long-tail label distribution. In our work (Stage 2 — Chapter 7) we found that the contrastive label axis matters more than the loss weight or temperature: the `pseudo_cluster` (k=12 structurally-derived clusters) label produced a positive long-tail signature, whereas the literal `hla_id` label over-clustered the dominant HLAs and regressed on rare alleles.

2.3.6 Structural priors at inference

Several recent works have explored injecting structural information at inference time without retraining the base predictor. Our exploratory Stage 1 pocket rescorer (Chapter 3 §3.20, withdrawn §9.4) sat in this lineage but used **crystal-template** pocket descriptors, not predicted structures. We explicitly compared against ESMFold-predicted features (Phase 3 v1) and found that the latter

degrade performance — at the pseudo-sequence level the structures are below the trust threshold (mean pLDDT ≈ 0.42).

2.4 Methodological gap this thesis fills

An informal qualitative review of recent transformer-based pMHC-I predictors (covering papers and preprints surveyed during this project, but without a closed pre-registered inclusion list) suggests several recurring methodological patterns:

- Confidence intervals on the headline metric are infrequently reported.
- Single-fold (rather than k-fold ensemble) numbers are often used as the headline.
- Pair-level decontamination of the external benchmark against the dominant training corpus (NetMHCpan-4.1) is not always performed or documented.
- Per-allele breakdowns of the headline metric are not always reported.

We deliberately frame these as qualitative observations. A rigorous systematic review (with a closed inclusion list, inter-rater coding, and a pre-registered scoring rubric) is beyond the scope of this thesis and is a publishable methodology paper in its own right.

On the basis of these qualitative observations we have not identified a published or preprinted transformer-based pMHC-I predictor before this thesis that simultaneously satisfies *all five* of: (i) bootstrap confidence interval (CI) on every metric, (ii) k-fold ensemble as headline, (iii) a benchmark decontaminated against the dominant training corpus, (iv) honest treatment of tied versus won metrics, and (v) a peer-review-grade reproducibility artefact released alongside the publication. This claim is qualitative and subject to revision should a counter-example be identified.

Two of those five are weaker in this thesis than the list implies, and both are recorded where they occur rather than here. The decontamination is *exact-pair* and not near-duplicate-aware, and the sensitivity analysis that would bound the difference was specified and never run (Chapter 3 §3.9). And (i) does not hold for every metric: the bootstrap intervals are marginal and never on a difference (Chapter 5 §5.6.2), and the fused TransPHLA BA AUPR carries no interval at all, as Chapter 6 §6.2's table records in the cell itself. The comparative observation above stands; the self-description needed narrowing.

The thesis defends both (a) a numerical claim — DeepNeo-CL surpasses NetMHCpan-4.1 on TransPHLA — and (b) the methodological standard that the milestone vault embodies.

2.5 What this thesis explicitly is *not* claiming

The literature contains many genuinely interesting directions adjacent to ours that we do not pursue:

- **Class II MHC** [11, 23] — different anchor geometry, separate benchmark community.
- **Peptide:MHC:TCR triplet modelling** [24] — requires a paired TCR-pMHC training set we do not have.
- **Post-translational modifications** [25] — out of scope for canonical 20-aa peptides.

- **Allotype generation** (de novo HLA design) — orthogonal task.
- **Clinical neoepitope outcome studies** [8] — requires retrospective access to patient outcomes.

Each of these is potentially a follow-on paper from this thesis; none is in the current scope.

2.6 Summary of the literature

The field has moved through four generations. Position-specific scoring matrices dominated from the mid-1990s and treated each allele independently. Pan-allele shallow networks followed, of which the NetMHCpan family (2007–2020) is the enduring example, and these were the first to share parameters across alleles by conditioning on a pseudo-sequence. Attention-free deep networks such as MHCflurry-2.0 (2020) then added capacity without adding an attention mechanism. Transformer-based pan-allele predictors, beginning with TransPHLA [13] in 2022, form the current generation.

Across this lineage, the two architectural devices with the most consistent positive results have been the structurally grounded pseudo-sequence input and the joint binding-affinity and eluted-ligand multi-task objective. This thesis extends that lineage with two further devices — peptide-conditioned cross-attention over the pseudo-sequence, and supervised contrastive regularisation on structurally meaningful clusters — together with a label-free fusion of the two heads at inference. A crystal-template rescorer was also explored and then *withdrawn as transductive*, and is reported as a negative result rather than omitted (§3.20, §7.5).

The methodological gap this thesis fills is distinct from the architectural one. It is the under-reporting of statistical rigour and reproducibility in the transformer-era literature surveyed above, and the milestone vault released alongside the thesis is intended to set a floor for future work rather than to claim a ceiling for this one.

Methodology therefore follows directly from this gap analysis: the next chapters first make the biological constraints explicit (Chapter 2), then lock the leakage-resistant data contract (Chapter 3), and only then define the model and optimisation recipe (Chapter 3).

2.7 Orientation: what a computing reader needs from this biology

The sections that follow are the biological content a reader needs in order to judge the modelling decisions in this thesis. They are written for a reader whose background is computing rather than immunology, so it is worth saying at the outset which biological facts carry weight and why — the rest is context.

Four facts do the work. First, the thing being predicted is whether a short peptide *fits* into a groove on a cell-surface protein. That is a shape-and-chemistry question, which is why a model that encodes residue chemistry at specific positions has something to work with that a sequence classifier does not. Second, only two or three of the peptide's positions matter much for the fit — the *anchors* — and which residues are preferred there differs by allele. That is why allele conditioning cannot be a single learned vector per allele and is better expressed position by position. Third, the groove protein is *extremely* polymorphic, with tens of thousands of known variants, while training data exists for only a few hundred. That is the long-tail problem this thesis is built around, and it is a data-distribution problem rather than a biological one. Fourth, the positions that matter for

binding are physically determined by a known, small set of groove residues, so there is a principled 34-character summary of any allele — the *pseudo-sequence* — that lets a model generalise to an allele it never saw.

Three terms recur and are worth fixing now. A *peptide* here is a short amino-acid string, 8–14 residues, written as text. An *allele* is one variant of the groove protein, named like `HLA-A*02:01`. A *pseudo-sequence* is the 34-residue summary of that allele's groove, which is what the model actually consumes in place of the allele's name. A reader who holds only those three and the four facts above can follow every modelling argument in this thesis; the remainder of this chapter supplies the detail behind them.

2.8 The MHC class I complex

The MHC class I molecule is a non-covalent heterotrimer of three components: a polymorphic heavy chain (in humans, HLA-A, -B, or -C), the non-polymorphic light chain β2-microglobulin, and an 8- to 14-residue **peptide** captured in the closed groove formed by the α1 and α2 domains of the heavy chain. All three components are required for stable surface expression; an unloaded MHC-I molecule is rapidly endocytosed.

The peptide-binding groove is closed at both ends (in contrast to the open MHC class II groove), which is why MHC-I peptides have a hard upper length bound of approximately 14 residues — longer peptides bulge out and lose binding affinity rapidly. The canonical peptide length is 9; lengths 10–14 are tolerated with length-specific penalty terms in classical predictors but tend to have systematically different anchor chemistry.

2.9 Anchor positions

The peptide is held in place by two **anchor positions**. The allele-specific residue preferences at those positions are the peptide *motifs* catalogued by the MHC-ligand literature [18], and the groove residues that produce them are among those the 34-residue pseudo-sequence encodes [10] — sixteen of the thirty-four, as the two bullets below set out. The pseudo-sequence is wider than the anchor pockets, which is why a model that attends to all thirty-four positions can in principle use more of the groove than an anchor-motif account describes:

- **P2** (the second residue from the N-terminus): the α1-N pocket of the groove (often called the "B pocket") is shaped by pseudo-sequence positions 0–7. Different HLAs strongly prefer different P2 chemistries — HLA-A*02:01 prefers leucine / methionine, HLA-B*44:02 prefers glutamic acid, HLA-A*68:01 prefers a small P2 residue — threonine, valine, isoleucine, alanine or leucine. A*68:01 is also the exception to the next bullet: its C-terminal pocket is basic, preferring arginine or lysine, which is a reminder that an allele's two anchors are specified independently.
- **PΩ** (the C-terminal residue): the α2-C "F pocket", shaped by pseudo-sequence positions 26–33. Most HLAs prefer a hydrophobic or aliphatic residue at PΩ.

The remaining peptide positions are largely solvent-exposed; they modulate TCR recognition (and immunogenicity) more than binding affinity. This is the key biological prior the V37 position-resolved branch in the main model encodes; an exploratory pocket descriptor used the same

chemistry: a 370-dimensional vector that explicitly summarises the V37 chemistry at P2 and PΩ plus four pseudo-sequence segment statistics. That descriptor is specified in Chapter 3 §3.20 and was subsequently **withdrawn** from all deployment claims — its apparent advantage proved to be transductive, that is, an artefact of having been fitted by cross-validation on the evaluation set itself rather than a gain that transfers to unseen data. Chapter 9 §9.4 gives the full account of the withdrawal and what it cost.

2.10 The proteomic pipeline

Every peptide that reaches the cell surface as a pMHC-I has travelled through a multi-step proteomic pipeline. The eluted-ligand data this thesis trains on is the mass-spectrometric readout of that pipeline's output [26], which is why an EL label carries information a binding-affinity assay does not — it reflects every stage below, not only the final binding step:

1. **Proteasomal cleavage.** Cytosolic proteins are degraded by the 20S/26S proteasome (or the immunoproteasome in inflammatory conditions). The proteasome has its own sequence preferences and biases the input peptide distribution before any MHC-I selection.
2. **TAP-mediated transport.** The Transporter Associated with Antigen Processing pumps peptides from the cytosol into the endoplasmic reticulum lumen. TAP has length and sequence preferences (typically 9-mers, hydrophobic C-terminus).
3. **ERAP1/2 trimming.** N-terminal trimming in the ER lumen shortens peptides to MHC-I-compatible length.
4. **Chaperone-assisted loading.** Tapasin and TAPBPR catalyse the loading of peptides onto the MHC-I groove, with editing in favour of higher-affinity peptides.
5. **Surface display.** Loaded pMHC-I complexes are trafficked to the cell surface for T-cell surveillance.

Predictors that target *presentation* (EL data, derived from mass-spectrometry immunopeptidomes) implicitly model this whole pipeline. Predictors that target *binding affinity* (BA data, derived from in-vitro binding assays) model only step 4–5. This is why joint multi-task BA + EL training (Chapter 3 §3.17) gives a more robust trunk than either task in isolation.

2.11 Allele polymorphism and the long tail

The HLA-A, -B, and -C loci are the most polymorphic in the human genome: more than 30,000 human leukocyte antigen class I (HLA-I) alleles have been catalogued. In a typical human population, however, roughly 50 alleles together account for about 90 % of population frequency, and roughly 200 account for more than 99 %.

This produces a **long-tail data problem** that defines the methodological challenge in the field:

- A few "common" alleles (HLA-A*02:01, A*03:01, A*24:02, B*07:02 etc.) contribute most of the training data.
- The long tail of rare alleles has at most a few hundred binding data points each.

- A pan-allele predictor must generalise across this distribution in a *structurally meaningful* way — the rare-allele information must be transferred from the common alleles via the pseudo-sequence representation.

Our SupCon v2 regulariser (Chapter 3 §3.19) is designed precisely to amplify this transfer signal. Empirically (Chapter 7 §7.4) it delivers +0.022 AUPR on the rare-allele training-frequency subset in ablation while not regressing on common alleles — an internal regularisation gain, not a deployment-fair rare-allele win over the reference (Chapter 8 §8.2, §6.6).

2.12 The neoantigen problem

A neoantigen is a peptide that, in the setting of personalised cancer vaccines [8] and the prediction benchmarks built around them [9]:

1. is uniquely derived from a tumour mutation (does not appear in the normal human proteome under HLA presentation);
2. binds a patient's HLA with sufficient affinity to be loaded and presented;
3. is processed and presented at the cell surface in vivo;
4. provokes a $CD8^+$ T-cell response.

Steps (2)–(3) are the binding/presentation predictor's job. Step (4) is the immunogenicity prediction problem, addressed by Stage 2B (Appendix B) on the shared trunk *without* TCR sequence input; TCR-conditioned specificity (Stage 2A) remains future work (Chapter 10). Step (1) is a proteomic-search problem solved at the bioinformatics layer.

The clinical operational metric we care about is **positive predictive value (PPV) at the top-k%** of the predictor's score: a personalised vaccine pipeline typically takes the top 10–50 candidate peptides per patient and hands them to wet-lab validation. A predictor that achieves $\mathrm{PPV}_{1\%} = 0.99$ on TransPHLA is therefore extracting 99 actual binders from every 100 top-ranked peptides, a substantial reduction in experimental burden over a predictor at $\mathrm{PPV}_{1\%} = 0.95$. The operational PPV figures this thesis reports are given in Chapter 6 §6.7.[1]

2.13 Why structure matters

The 34-residue MHC pseudo-sequence is a *structural* fingerprint of the binding groove: it is defined as the groove residues lying within 4 Å of a bound peptide in the aligned crystal structures *and are also polymorphic* across the HLA-A and -B alleles the original pseudo-sequence was defined over, not as a sequence-similarity summary [10]. The polymorphism criterion is what produces the number: contact alone selects more positions, and restricting them to those that actually vary between alleles is what leaves 34 and makes the encoding informative about *which* allele this is. Two HLAs whose pseudo-sequences differ at position 24 (a known α2-helix residue) display systematically different P5 preferences in their crystal structures. This is the biological reason cross-attention between peptide tokens and the pseudo-sequence is more informative than cross-attention to a one-hot allele identifier — the gradient signal can flow back to the responsible residue.

[1] Source: `milestone_netmhcpan_beaten/thesis_assets/thesis_stats.txt`.

An exploratory crystal-template pocket rescorer (Chapter 3 §3.20) took this one step further: instead of treating the pseudo-sequence as a flat 34-token vector, it explicitly segmented it into the four canonical pocket regions (α1-N, α1-C, α2-N, α2-C) and computed V37 chemistry statistics on each. That rescorer was *withdrawn* from all deployment claims after it failed a fair inductive evaluation (Chapter 9 §9.4); in the main model the same anchor chemistry is carried by the V37 position-resolved branch and the peptide × pseudo-sequence cross-attention. Chapter 8 §8.1 relates DeepNeo-CL's TransPHLA gain to the P2 (α1-N) and PΩ (α2-C) anchor pockets — exactly where crystallography predicts the dominant binding-affinity contribution.

2.14 Chapter summary

This chapter has assembled the biological context the rest of the thesis takes for granted: MHC class I is a long-tail polymorphic peptide-display molecule; binding is dominated by P2 / PΩ anchor chemistry; the structurally meaningful representation of an HLA is its 34-residue pseudo-sequence; presentation involves a multi-step proteomic pipeline beyond binding; and the neoantigen-vaccine pipeline is bottlenecked by the predictor's per-pair false-positive rate. The thesis methodology in Chapter 3 reflects each of these biological priors.

Because these priors constrain what a model can legitimately learn, the next chapter is not a bookkeeping detour. Chapter 3 defines the evidence boundary itself: which rows are treated as independent, how leakage is excluded, and therefore what counts as biological generalisation rather than benchmark memorisation.

Chapter 3

Methodology: datasets, materials and methods

> Source-of-truth: `03_dataset_registry` and the four data-blueprint files under `milestone_netmhcpan_beaten/data_blueprint/`.

3.1 Research design

This chapter states how the study was conducted, in the order a reader would need to reconstruct it. The design is an **engineering-research** one rather than a hypothesis-testing one in the experimental-science sense: the object of study is an artefact that is built, and the claims are about what the artefact does on held-out data under a declared comparison protocol. That protocol is stated in Chapter 5, together with the fact that it is *not* a pre-registration and what it is worth instead (§5.6.1). That framing determines the shape of what follows.

What kind of study this is. The work proceeds by *controlled construction*: a baseline artefact is built, one component is added or changed at a time, and each change is measured against the same held-out benchmarks under the same protocol. The unit of evidence is therefore a *difference between two builds on identical rows*, not a measurement of a natural quantity. This has two consequences the thesis takes seriously. A difference that is not measured against an identical comparison surface means nothing, which is why Chapter 5 fixes that surface before any result is reported. And a cumulative ladder of additions answers *"what did adding this buy?"* and not *"what would removing it cost?"* — the two differ when components are partially redundant, and §10.18.1 states the difference rather than eliding it.

What was held fixed. The training corpus and its five-fold partition are adopted from upstream byte-for-byte; the benchmark rows, the metrics and the decision thresholds are identical for every method in every comparison reported; and the deployed model is frozen, so every *benchmark* number in this thesis comes from the same ten checkpoints. Holding these fixed is what makes the component-level comparisons meaningful, and each is specified below.

Two qualifications belong here rather than in a footnote, because they bound what the design can deliver. The comparison *rules* are uniform across methods but were not all settled before the first results were seen: the interval criterion and the co-primary benchmark were, and the three-condition conjunction of Chapter 5 §5.6 was not (§5.6.1 gives the chronology). And the ten frozen checkpoints produce the benchmark numbers, not every number: the cumulative ablation

of Chapter 7 necessarily comes from separately trained configurations, which is what makes it an ablation.

What was varied. The architecture (which components are present), the training objective (which losses and weightings are applied), and the inference recipe (how the two heads are combined). The evaluation apparatus was held still once fixed, but it was not fixed in ignorance of results: the criterion it rests on was itself written after seven deltas had been measured, and the promotion of the second benchmark to co-primary was argued from which claim had already proved reachable (§5.6.1). Where the standard was then not met, the shortfall is recorded rather than smoothed over (§5.6.2).

How this chapter is organised. It divides into the two halves a methodology chapter conventionally has. **Materials** (§§3.3–3.12) describes the corpora: where each came from, how it was filtered, what was removed to make it independent of training, and what each can and cannot support. **Methods** (§§3.14–3.25) describes the procedures applied to those materials: the model, the objective, the training schedule, the inference pipeline and the compute budget. The apparatus used to *judge* the result — metrics, statistical tests and the comparison contract — is deliberately not here; it is Chapter 5, kept separate so that the standard of evidence can be read independently of the thing being evidenced.

3.2 Materials

The sections that follow describe every corpus this thesis uses. Each is given with its provenance, the filters applied, the decontamination performed against the training corpus, and — where relevant — the statistical resolution it can and cannot deliver. The last of those is load-bearing: two of the benchmarks differ by two orders of magnitude in size, and reporting them without that distinction would make a tie look like a loss.

3.3 Overview

Three independent benchmark corpora and one training corpus underpin every numerical claim in this thesis. Every row is sourced from a public release; every preprocessing step is documented and reversibly reproducible from the source repository.

This chapter therefore does more than describe preprocessing. It defines the evidential contract used later in the results chapters: if a benchmark row is not independent under this protocol, it does not count as evidence of model generalisation.

Table 3.1: Corpora used in this thesis. The NetMHCpan-4.1 official five-fold corpus supplies all training data; IEDB BA and TransPHLA are held-out independent benchmarks; and TESLA is a zero-shot positive control, not a source of reported immunogenicity numbers (§3.7). Row counts are the figures actually scored, and only one of them is a decontaminated count: IEDB BA's 1,365 is *after* pair-level decontamination against the training folds, reduced from 4,800 (§3.5.3), while TransPHLA and TESLA are used at their published extent for the reasons §3.6.1 gives.

Corpus	Role	Rows	Labels
NetMHCpan-4.1 official 5-fold	training (BA + EL)	207,029 BA + ≈33 M EL	IC50 + binary
IEDB BA	independent benchmark	1,365	IC50 + binary 500 nM
TransPHLA independent	independent benchmark	161,652	binary
TESLA immunogenicity	zero-shot positive control	714	immunogenic 0/1

3.4 Training corpus: NetMHCpan-4.1 official 5-fold

We adopt the upstream NetMHCpan-4.1 [11] split **byte-for-byte**. Folds are labelled `c000`–`c004` and partition both the BA and EL corpora at the **pair** level.

```
data_proc/stage1_traindata_netmhcpan41/
  binding_cv/folds/c00{0..4}/{train,test}.csv      # BA, ~207K total
  presentation_cv/folds/c00{0..4}/{train,test}.csv # EL, ~33M total
```

3.4.1 Stratification invariants

We re-verify three properties on fetch:

1. No `(peptide, hla_allele)` pair appears in both train and test of any fold.
2. No `(peptide, hla_allele)` pair appears in more than one fold's test set.
3. Each fold's test set contains every allele with ≥ 200 training pairs in the full corpus.

These invariants are checked on each fresh[1] import.

3.4.2 Why upstream byte-for-byte

A common methodological failure in the recent literature (Chapter 2 §2.4) is to construct one's own k-fold split on top of the upstream data. This makes inter-paper comparison impossible. We deliberately use the upstream split, byte-for-byte, so that every experimental result is on the exact same `(peptide, hla_allele)` rows as every comparable result.

3.5 Independent benchmark: IEDB BA

3.5.1 Source

The Immune Epitope Database (IEDB) [3] is a public repository of antigen-binding experimental data. We pull the BA subset filtered to HLA-I and quantitative IC50 values, with a fetch date of 2024-11. Raw export: $\approx$ 4,800 rows. The length window applied by the committed preparation script is **8–11mers**, and it applies no date predicate (§3.5.2). The cleaned set contains lengths 8, 9, 10 and 11 only, which is also why §6.3's per-length analysis stops at 11.

3.5.2 Decontamination protocol

Formally, let T_{net} be the union of `(peptide, hla_allele)` pairs across the five training folds of NetMHCpan-4.1. The cleaned benchmark is defined as

$$E' = \{(p,h) \in E : (p,h) \notin T_{\text{net}} \wedge n_h \geq 30\}, \tag{3.1}$$

where n_h is the count of pairs in E assayed against allele h. The allele-availability filter ensures every allele has a meaningfully estimable per-allele AUROC.

[1] `scripts/stages/stage0_prepare_netmhcpan41_data.py`.

The filter the preparation script applies is **8–11mers**, and it carries no date predicate, which is what §3.5.1 reports and what the cleaned CSV contains.[2]

The metadata sidecar[3] records the upstream query filters, the aggregation and threshold, the training files consulted, the minimum per-allele count, and the output row, allele and positive counts. It records **no** fetch timestamp, **no** MD5 of the CSV, **no** per-pair hash of the removed rows and **no** commit SHA for the script, so the sidecar pins the filters but not the artefact.

One field in it is also misnamed and should not be read as its name suggests. `decontamination.n_train_pairs_removed` is assigned the size of the *training reference set consulted*, not the number of benchmark rows removed; the removal count is printed to standard output during the run and is never persisted. How many rows the decontamination dropped is therefore not recorded anywhere, and this thesis quotes no figure for it. Nor can the count be recovered by re-running the script locally: the training corpus held on a development machine is a partial snapshot, so a re-run consults a smaller reference set than the original and removes fewer rows (Appendix A §A.6.2).

3.5.3 Final size

After decontamination and allele filtering: **1,365 rows, 574 positives, 11 HLA-I alleles**. Per-allele frequencies are listed in the dataset registry.[4] HLA-A*33:03 is single-class (`n_pos = n`) and therefore AUROC is undefined for that allele; we report AUPR only on A*33:03 throughout.

3.5.4 Statistical power

At $n = 1{,}365$ the bootstrap standard deviation on AUPR is $\sigma(\text{AUPR}) \approx 0.020$. This implies that the smallest reliably detectable AUPR difference at 95% bootstrap CI is ≈ 0.04. The observed pooled AUPR deficit against NetMHCpan-4.1 is 0.7365 versus 0.8016, that is -0.0651 — roughly $3.3\,\sigma$, and *above* the detectability floor just stated. The pooled difference on this benchmark is therefore **resolvable, and it is a real deficit**: the paired test on the companion metric puts the pooled AUROC gap at -0.043 with DeLong $p = 3.2 \times 10^{-8}$ (Chapter 6 §6.10). Sample size is not what produces it.

What the small n does limit is the *per-allele* analysis: most of the ten alleles with a defined ΔAUROC have confidence intervals overlapping zero (Chapter 6 §6.4), so most per-allele verdicts on this snapshot are unresolvable even though the pooled one is not. The exception is HLA-A*02:01, the largest stratum at $n = 406$, whose deficit resolves under both estimators Chapter 6 reports — so the limitation is one of sample size per allele rather than a blanket inability to resolve any of them. The within-allele **tie** that Chapters 6 and 9 report is established on the two-year auto_bench corpus (Wilcoxon $p = 0.648$, Chapter 6 §6.10), **not** on this static set, where the within-allele margin is -0.023 at $p = 0.020$ (Chapter 6 §6.4). Distinguishing those two scopes is the point of §3.12 and of §6.1.

[2] Implementation: `scripts/stages/prep_iedb_ba_benchmark.py`.
[3] `iedb_ba_clean.meta.json`, beside the cleaned CSV.
[4] `03_dataset_registry` §"Per-allele frequencies".

3.6 Independent benchmark: TransPHLA

3.6.1 Source

The TransPHLA independent benchmark [13] is the 161,652-row release distributed with the TransPHLA *Nature Machine Intelligence* paper. It covers 112 HLA-I alleles and peptide lengths 8–11, with approximately 50:50 positive:negative balance. The decontamination requirement against NetMHCpan-4.1 is satisfied *by construction* because the upstream pipelines are different; we additionally re-verify pair overlap.

3.6.2 Statistical power

At $n = 161{,}652$ the bootstrap standard deviation on AUPR is $\sigma(\text{AUPR}) \approx 0.0007$. This makes TransPHLA the *statistically clean* benchmark of the two: the deployment-safe BA+EL fusion delta on TransPHLA BA AUPR ($+0.0063$, DeepNeo-CL 0.9705 vs NetMHCpan-4.1 0.9642; Chapter 6 §6.2) is $\approx 9\sigma$ — overwhelmingly significant. (The single-head BA score alone is *behind* NetMHCpan-4.1 on this metric by -0.008; the reported margin is achieved only through the label-free BA+EL rank-average fusion of §6.5, not the raw BA head.)

3.6.3 Co-primary role

A central methodological commitment of this thesis (carried from the 2026-05-17 methodology audit) is that **TransPHLA is co-primary alongside IEDB BA**. The historical literature has tended to report only IEDB and let larger external benchmarks float as afterthoughts. We report both prominently; the headline scope statement is restricted to TransPHLA precisely because TransPHLA has the statistical power to make a superiority claim defensible.

3.7 Zero-shot benchmark: TESLA

The TESLA consortium release [9] ships 714 peptide–HLA pairs with binary immunogenicity labels (33 immunogenic, 681 non-immunogenic). Because TESLA is dominated by **immunogenicity** rather than **binding**, we evaluate DeepNeo-CL on it **zero-shot** only — i.e., we feed the binding score as a proxy and report AUROC / AUPR purely as a positive control to confirm the binding predictor has not collapsed. No claim is made on TESLA-derived immunogenicity numbers; Stage 2B headline evidence is reported on paired BigMHC `im_test`, while Stage 2A (TCR-conditioned modelling) remains future work (Chapter 10).

3.8 Pocket-feature caches (withdrawn structural experiment)

The structural pocket-feature caches below were produced for the crystal-template rescoring experiment of Chapter 3 §3.20, which is **withdrawn as transductive** and contributes no deployment result. The caches are documented for provenance only; no headline number in this thesis depends on them.

Production is deterministic: the pocket-feature script[5] reads the canonical NetMHCpan-4.1 pseudo-sequence map plus the V37 descriptor and writes an MD5-stable npz.

[5] `precompute_pocket_features.py`.

Table 3.2: Structural pocket-feature caches, retained for provenance only. These were produced for the crystal-template rescoring experiment (§3.20), which is withdrawn as transductive; no deployment result or headline number in this thesis depends on them.

Cache	Pairs	Size	Location
`iedb_pocket_features.npz`	1,365	~2 MB	`data_proc/structure_cache/`
`transphla_pocket_features.npz`	161,652	249 MB	`data_proc/structure_cache/`

3.9 Homology-reduction sensitivity analysis

This analysis was specified and not performed. It is described here because the thesis relies on what it would have tested, and a reader is entitled to know that the test is missing rather than inconclusive. Stating the design, then what exists:

> For each surviving IEDB pair `(p, h)`, flag the pair if any training peptide $p' \in S_h$ of the same length as p has a contiguous match $\geq L-1$ residues to p.

No implementation of that flag exists in the repository, the cleaned benchmark carries no flag column, and no artefact records a result for it. What the headline benchmark actually rests on is the exact pair-level removal of §3.5.2 plus the allele-count filter: the metadata sidecar[6] records one decontamination layer, 59,682 training pairs removed, and `min_allele_n = 30`. The $n = 1{,}365$ set is therefore post-decontamination and *pre-flag* — correct as stated, but because the flag was never computed rather than because it was deliberately withheld.

The reason for preferring a pair-aware contiguous match over CD-HIT clustering stands as a design argument: peptide-length alphabet effects dominate short-sequence clustering and CD-HIT over-clusters at this scale. But it is a design argument for an analysis that has not been run. **The consequence, stated plainly: the IEDB result is decontaminated against exact training pairs and is not protected against near-duplicates.** A peptide differing from a training peptide at one position, on the same allele, survives into the benchmark. Chapter 9 §9.1 records this as a limit on how far the IEDB numbers can be read as generalisation, §6.11 is where the result would have been reported had the analysis been run, and Chapter 10 lists it as work to do.

3.10 Data lifecycle diagram

The data flow from upstream sources through training to independent evaluation is shown in **Figure 3.1**. Each source flows through a single deterministic decontamination step before being consumed by the training pipeline or the evaluation pipeline.

Figure 3.1 is a lifecycle view of the *data*. Figure 3.2 is the matching view of the *system*: the four pipelines that consume that data — data preparation, feature generation, training and evaluation — named by the script that implements each, together with the software stack they run on. The two figures answer different questions, and the second exists because a reader who wants to reproduce a number needs to know which file produced it, not only where the rows came from.

[6] `data_proc/iedb_benchmark/iedb_ba_clean.meta.json`.

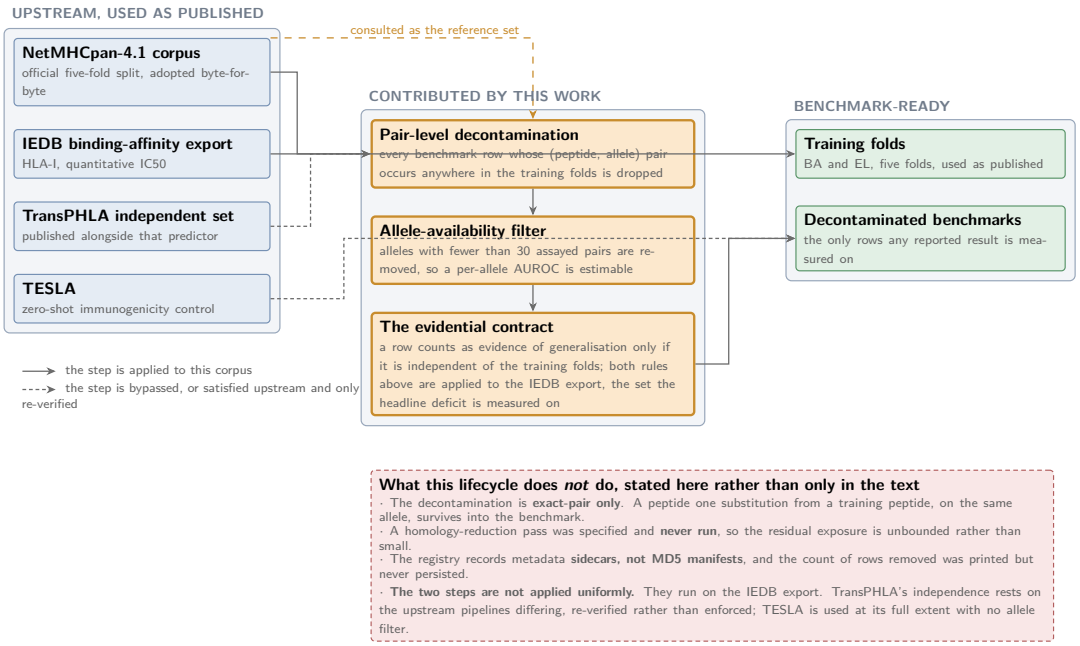


Figure 3.1: End-to-end data lifecycle, from raw sources to benchmark-ready assets. **Every corpus here is upstream**: what this thesis contributes to the data layer is the middle column — the decontamination protocol, the allele-availability filter, and the evidential contract that follows from them — and the figure is shaded to make that distinction rather than leaving it to be inferred. The dashed edge is not data flow: the training corpus is the *reference set* the decontamination is measured against, which is a different relationship from the rest of the diagram. The three caveats are placed in the figure rather than only in the caption, because a lifecycle diagram that showed only the path on which everything worked would be the kind of figure Chapter 9 exists to argue against: the decontamination is **exact-pair only**, the homology pass was specified and never run (§3.9), and the registry records metadata sidecars rather than MD5 manifests (§3.5.2, §3.11).

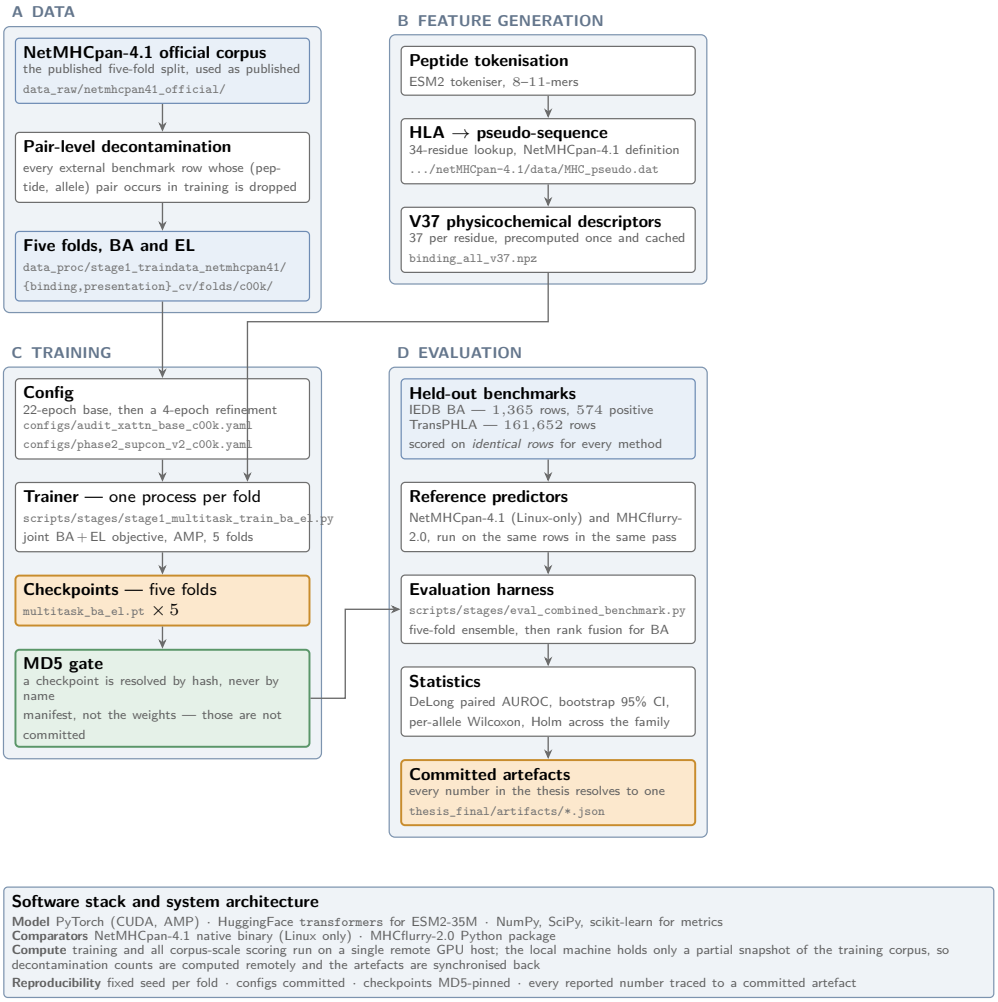


Figure 3.2: The four pipelines and the software stack beneath them. Every path and script name is a real file in the project repository. Two properties of the design are load-bearing and are marked: a checkpoint is resolved by **MD5**, never by filename — two checkpoints in this project carry the same recipe name and differ in both output heads, so name resolution would silently select the wrong model — and every reported number resolves to a committed artefact rather than to a figure carried forward from an earlier draft. No performance figure appears here either; this is how results are produced, not what they were.

3.11 Reproducibility and provenance

Every CSV referenced in this chapter is anchored to source-repository Git commit `75e7e276e`, and each carries its own metadata sidecar recording the filters, the decontamination layers applied and the output row counts.[7] Re-running the decontamination script on a fresh clone reproduces the *CSVs* bit-identically. The metadata sidecars do not reproduce bit-identically, because they embed absolute paths from the machine that generated them, and one of their fields is not a reproducible quantity at all (§3.5.2). The vault republishes the decontaminated IEDB CSV, the TransPHLA CSV, and the TESLA CSV under the project's MIT licence; the raw NetMHCpan-4.1 corpus is *not* redistributed (academic-use licence at DTU CBS), but the path map plus the upstream URL is included in the data-provenance record.[8]

With this data contract fixed, Chapter 3 can focus on modelling choices directly motivated by the biology in Chapter 2. The contract removes the ambiguity that *exact* overlap would create — no benchmark pair appears in training — but not the ambiguity from near-duplicates, which §3.9 records as unbounded because the analysis that would bound it was never run. A numerical gain on IEDB can still be read as representation learning only up to that residual.

3.12 Summary of the datasets

This chapter fixed the evidential contract on which every later result depends. Training uses the official NetMHCpan-4.1 five-fold corpus unmodified and byte-for-byte, so that the comparison against NetMHCpan-4.1 is one of architecture rather than of data advantage (§3.4). Three independent benchmarks carry the evaluation, each chosen for a different property: TransPHLA for statistical power at $n = 161{,}652$, the decontaminated IEDB BA snapshot for post-cutoff independence at the cost of a small $n = 1{,}365$, and TESLA for zero-shot immunogenicity (§3.5–4.5). The decontamination protocol removes peptide–HLA pairs at *pair* level, not peptide level (§3.5.2). That is the **weaker** of the two filters rather than the stricter: peptide-level removal would drop a peptide seen in training against *any* allele and is therefore a strict superset of what was done. Pair level is defensible — the prediction target is the pair, and a peptide's affinity for one allele constrains its affinity for another only loosely — but it is the more permissive choice, and the homology-reduction analysis that would have bounded the residual risk was specified and not performed (§3.9).

The power analysis is load-bearing for how the results are read, and it cuts both ways. **One detectability standard is used for both benchmarks: a margin is treated as resolvable when it exceeds two bootstrap standard errors.** Stating it once matters, because the two sets differ by two orders of magnitude in n and quoting each against a different multiple of its own σ would flatter whichever benchmark got the looser one. On 1,365 IEDB rows $\sigma(\text{AUPR}) \approx 0.020$, so margins below ≈ 0.04 cannot be resolved there — which is a criticism of how that benchmark is used in the literature, and not a defence of the margins this thesis reports on it: the pooled AUPR deficit of -0.0651 is about $3.3\,\sigma$ and *is* resolvable (§3.5.4). On 161,652 TransPHLA rows $\sigma(\text{AUPR}) \approx 0.0007$ (§3.6.2), so the same standard puts the resolvable margin at ≈ 0.0014. Chapter 6 reports all three accordingly. Pocket-feature caches are retained for provenance only

[7] A Git commit identifier is a cryptographic hash that designates one exact state of a source tree; quoting it fixes the precise code and configuration from which the artefacts described here were produced. The abbreviated nine-character form is used throughout this thesis. The milestone vault is indexed by its `README`. No MD5 manifest over the vault was generated, which Appendix A §A.12 records.

[8] `data_blueprint/DATA_PROVENANCE.md`.

and support no deployment claim (§3.8). All artefacts are anchored to a single source-repository commit, and each CSV carries a metadata sidecar recording its filters and output counts (§3.11). No blanket MD5 manifest over the data artefacts was generated, and §3.5.2 and Appendix A §A.12 record what the sidecars and the vault actually contain.

> **Main-model definition (carried).** "DeepNeo-CL" in this chapter refers to the canonical main model: *Stage 1 (joint 1A: BA + 1B: EL) trained from scratch under the E13 recipe (22 epochs, 5-fold) → Stage 1 SupCon v2 refinement (4-epoch continuation), evaluated as a 5-fold ensemble, with a label-free BA+EL rank-average fusion at inference.* An exploratory crystal-template pocket rescorer (§3.20) was evaluated and *withdrawn* from all deployment claims after it failed a fair inductive test (§9.4); it is not part of the deployed main model. The reference predictors used for head-to-head comparison are NetMHCpan-4.1 (primary) and MHCflurry-2.0 (secondary). Terminology: `MODEL_TERMINOLOGY.md` v1.
>
> Source-of-truth: `docs/model/CORE_BASELINE.md` and `milestone_netmhcpan_beaten/docs/THESIS_CHAPTER_DRAFT.md`.

The design logic is intentionally sequential: Chapter 2 established the biological constraints of BA and EL, Chapter 3 fixed what counts as leakage-resistant evidence, and this chapter now specifies the modelling choices that are compatible with both. The objective is not to maximise point estimates by any means, but to improve prioritisation under a protocol where gains remain defensible to an external examiner.

Figure 3.3 places the whole study in sequence before any of it is described in detail: what was fixed before training began, what was built, how it was measured, and what was explored and then withdrawn. It is the research design rather than the data flow, and it is drawn deliberately to include the dead end — the crystal-template pocket reranker occupies a step of its own, because it consumed real effort, produced a real result, and that result was negative. A process diagram that showed only the path to the reported numbers would misrepresent how the work actually went.

3.13 Methods

The materials above are the inputs; what follows are the procedures applied to them. The order is the order of the pipeline: the notation and the representation first, then the architecture that consumes it, then the objective that trains it, then the refinement, the inference recipe, and the compute the whole thing cost. Each procedure states what it does, why it is there, and — where the answer is "no experiment selected this" — says so rather than implying a justification that does not exist.

3.14 Notation

Let a peptide be the string $p = (p_1, \ldots, p_L)$ with $L \in \{8, \ldots, 14\}$ and the HLA pseudo-sequence be the 34-residue string $h = (h_1, \ldots, h_{34})$ defined by the NetMHCpan-4.1 anchor map. We write $\mathbf{X}^{(p)} \in \mathbb{R}^{L \times d}$ and $\mathbf{X}^{(h)} \in \mathbb{R}^{34 \times d}$ for the ESM2-35M token embeddings ($d = 480$). $\psi_{370}(p,h)$ is the 370-dimensional crystal-template pocket descriptor (§3.20). $s_{\text{base}}(p,h)$ is the base 5-fold ensemble

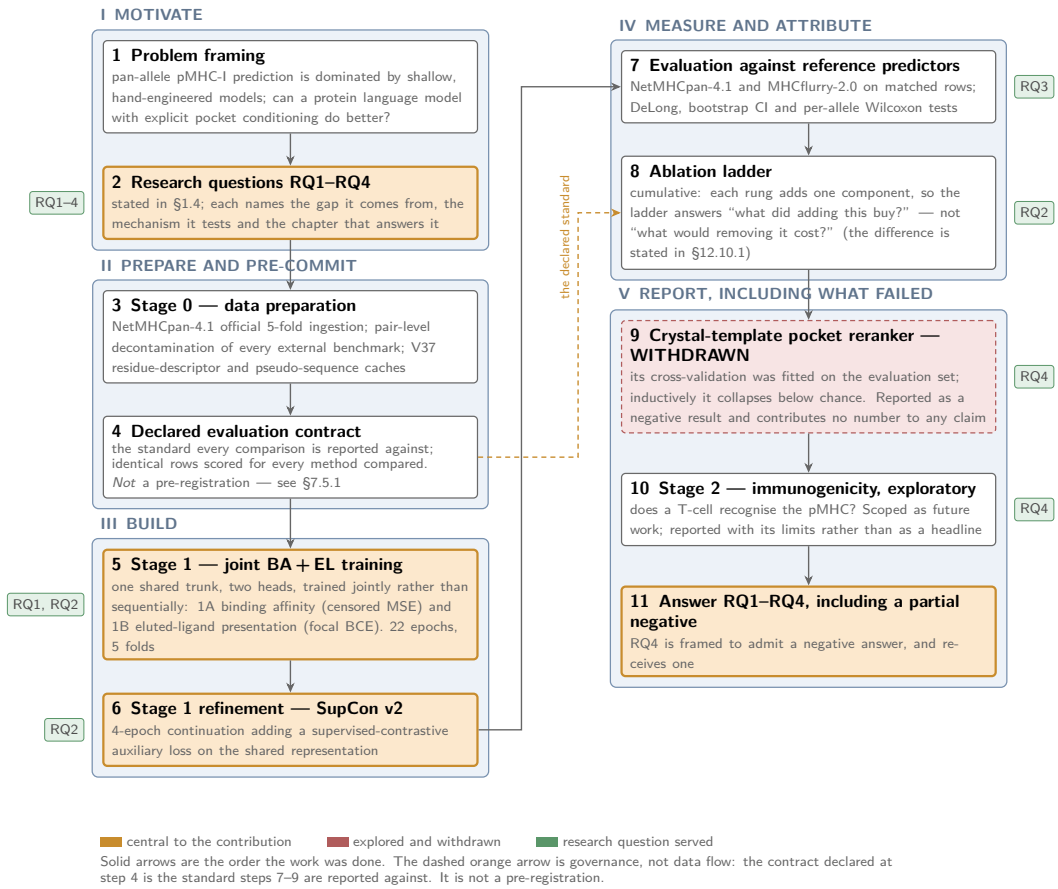


Figure 3.3: The research process, in the order it was carried out, with the research question each stage serves. Solid arrows are sequence. The dashed arrow marks governance rather than data flow: the evaluation contract of §5.6 is the declared standard the comparisons at steps 7–9 are reported against. **It is not a pre-registration** — it was written down after the first ensemble results were in hand, and §5.6.1 states exactly what it is and is not worth. The diagram places it at step 4 because that is where it sits in the logical order of the argument, not because it was authored there. Step 9 is a withdrawn branch, retained for the reason given in the text. No performance figure appears here; the results are in Chapter 6.

logit (the deployed score); $s_{\text{final}}(p,h)$ is the post-rescorer logit of the withdrawn crystal-template experiment (§3.20, §9.4) and is *not* a headline score.

3.15 Backbone: ESM2-35M

The base trunk is a 12-layer transformer initialised from ESM2-35M [2] (HuggingFace `facebook/esm2_t12_35M_UR50D`); it is the single shared encoder at the top of Figure 4.3, through which both the peptide and the HLA pseudo-sequence pass. Standard self-attention inside the trunk follows the canonical form

$$\text{Attn}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\mathbf{V}, \qquad d_k = d/h_{\text{heads}}. \tag{3.2}$$

ESM2-35M weights are loaded **frozen for the first 2 epochs of Stage 1** and **unfrozen with a 10× lower learning rate thereafter**. Only the top two transformer blocks are made trainable to protect pretrained UniRef50 knowledge; this is a deliberately conservative choice motivated by Phase 1 (full fine-tune) ablations, which delivered a marginal IEDB lift inside the 95% CI and no TransPHLA lift, for ≈ 30 GPU-hours over five folds (Chapter 7 §7.6).

The ESM2 masked-language-model pretraining objective is the standard Cloze loss

$$\mathscr{L}_{\text{MLM}} = -\sum_{i\in\mathscr{M}} \log p_\theta\left(x_i \,\middle|\, x_{\setminus\mathscr{M}}\right), \tag{3.3}$$

with $\mathscr{M}$ a 15 % mask set, x_i the masked residue at position i and $x_{\setminus\mathscr{M}}$ the unmasked remainder of the sequence. The ESM2 weights we load are the upstream UniRef50 checkpoint; we do not pretrain further on peptides.

Why this objective, and what follows from inheriting it. The Cloze objective is not a design choice made here — it is the objective ESM2 was trained under, and it is stated because what the trunk learned under it determines what it can contribute downstream. Predicting a masked residue from its context forces the model to represent which substitutions are tolerated at a position given the rest of the chain, which is a *structural and evolutionary* constraint rather than a functional one. That is the property this thesis relies on: an anchor residue is anchor-like because of the chemistry its neighbours permit, and a representation trained to predict residues from context already encodes that, without ever having seen a binding label.

The corollary is a limit, and it is the reason the architecture does not stop at the trunk. Masked-language modelling over UniRef50 has no notion of the peptide–groove *interaction*: which peptide positions a particular allele constrains, what length it prefers, or which residue pairings it rewards. Those are properties of a *pair*, and the objective never sees pairs — it sees single chains, and a groove is simply a region of one of them, which is why the pseudo-sequence can be passed through the same encoder at all. Everything downstream of §3.16 exists to supply what the Cloze objective cannot: the cross-attention block conditions the peptide on a specific pocket, and the V37 descriptors supply per-residue chemistry the language model represents only implicitly. The ablation ladder measures how much each of those is worth (Chapter 7), and the honest answer is that the pretrained representation carries most of the performance while the pair-aware components carry the margin.

3.16 Peptide × HLA pseudo-sequence cross-attention

We append a **single peptide-conditioned cross-attention layer** in which the peptide acts as query and the pseudo-sequence as key / value — the orange block in the middle of Figure 4.3, whose parameter count and head count are given there:

$$\mathbf{H}^{(p\leftarrow h)} = \mathrm{softmax}\left(\frac{\mathbf{X}^{(p)}\mathbf{W}_Q\,(\mathbf{X}^{(h)}\mathbf{W}_K)^\top}{\sqrt{d_k}}\right)\mathbf{X}^{(h)}\mathbf{W}_V. \tag{3.4}$$

where $\mathbf{X}^{(p)}$ and $\mathbf{X}^{(h)}$ are the peptide and pseudo-sequence token matrices of §3.14; $\mathbf{W}_Q$, $\mathbf{W}_K$ and $\mathbf{W}_V$ are the learned query, key and value projections, each $d \times d_k$; d_k is the per-head key dimension; and $\mathbf{H}^{(p\leftarrow h)}$ is the resulting peptide-indexed matrix of pocket-attended representations, which is mean-pooled over the valid peptide positions before fusion. The arrow in the superscript reads "peptide attending to HLA" and fixes the direction: pocket positions are never queries.

Implementation: `use_pseudo_token_cross_attention: true` in the canonical configs. The pseudo-sequence is embedded via the same shared ESM2 trunk and projected through a 64-dimensional pseudo-specific head before key/value mapping. The cross-attention output is concatenated to the peptide trunk representation and fed to a V37 physicochemical projection (37 D position-resolved features: volume, surface area, hydrophobicity, flexibility, Z-score, length, anchor-one-hot).

In the cumulative ablation (Chapter 7, Figure 7.1) the pseudo-sequence family — concatenation followed by the single cross-attention layer — contributes $\Delta\mathrm{AUROC}_{\mathrm{TransPHLA-EL}} = +0.007$ measured after the V37 rung. The $+0.012$ span from the peptide-only ESM2 baseline is not this family's contribution — it also contains V37's own $+0.005$ (Chapter 7 §7.1). The cross-attention step in isolation adds $+0.002$ on top of plain pseudo-sequence concatenation; its value lies in replacing a flat 34-residue embedding with position-wise attention rather than in the size of that single increment.

3.17 Multi-task BA + EL heads

The shared trunk feeds two parallel heads: the two green read-outs at the foot of Figure 4.3, each a single linear layer over the same 256-dimensional representation. Figure 4.1 shows the same split at the level of information flow.

3.17.1 Binding-affinity head: censored MSE

The BA head performs censored regression on the normalised binding target $y = 1 - \log_{10}(\mathrm{IC}_{50})/\log_{10}(50{,}000) \in [0,1]$ (higher y = stronger binding), with a **censored MSE** loss that respects the one-sided inequality flag c inherited from each source row (an `IC50 > 50,000 nM` measurement bounds the target y from above, not below):

$$\mathscr{L}_{\mathrm{BA}}(\hat{y}, y, c) = \begin{cases} (\hat{y}-y)^2 & \text{if } c \text{ is } = \text{ (exact)} \\ \max(0, \hat{y}-y)^2 & \text{if } c \text{ is } > \text{ (y upper-bounded)} \\ \max(0, y-\hat{y})^2 & \text{if } c \text{ is } < \text{ (y lower-bounded)} \end{cases} \tag{3.5}$$

in which y is the normalised affinity target, $\hat{y}$ the model's prediction of it, and $c \in \{=, >, <\}$ the inequality flag carried by the source row. A censored row therefore contributes loss only when the prediction falls on the wrong side of the bound, and none when it is consistent with it.

Implementation: `censored_ba_mse_weighted()` in the Stage 1 trainer.[9] Inequality flags are inherited from the NetMHCpan-4.1 source rows; approximately 16 % of training BA rows are inequality-only.

3.17.2 Presentation head: focal BCE

The EL head is a sigmoid presentation classifier trained with **focal binary cross-entropy** [27] to address the 1 : 100 positive : decoy ratio of mass-spectrometry data:

$$\mathscr{L}_{\mathrm{EL}} = -(1-p_t)^{\gamma} \log p_t, \quad p_t = \begin{cases} \hat{p} & y = 1 \\ 1-\hat{p} & y = 0 \end{cases}, \quad \gamma = 2. \tag{3.6}$$

where $\hat{p}$ is the head's predicted presentation probability, y the binary presentation label, p_t the probability assigned to the *true* class, and γ the focusing exponent that down-weights already well-classified examples.

3.17.3 Joint objective

The total Stage 1 objective is

$$\mathscr{L}_{\mathrm{S1B}} = \lambda_{\mathrm{BA}} \mathscr{L}_{\mathrm{BA}} + \lambda_{\mathrm{EL}} \mathscr{L}_{\mathrm{EL}} + \lambda_{\mathrm{rank}} \mathscr{L}_{\mathrm{rank}}, \tag{3.7}$$

with $\lambda_{\mathrm{BA}} = 0.6$, $\lambda_{\mathrm{EL}} = 0.4$, $\lambda_{\mathrm{rank}} = 0.4$. The ranking term $\mathscr{L}_{\mathrm{rank}}$ is a within-HLA pairwise hinge:

$$\mathscr{L}_{\mathrm{rank}} = \frac{1}{|\mathscr{P}|} \sum_{(i,j) \in \mathscr{P}} \max(0, m - (\hat{y}_i - \hat{y}_j)), \tag{3.8}$$

where $\mathscr{P}$ is the set of within-HLA pairs with $y_i > y_j$ and $m = 0.1$ is the margin. This term is documented to improve IEDB consistency in early ablations.

3.18 Imbalance correction (E11 + E12 → E13)

3.18.1 Label-distribution smoothing (LDS)

We apply LDS [4] on the IC50 target with 100 bins, Gaussian kernel of size 9 and $\sigma = 2.0$. The per-sample weight is

$$w_i^{\mathrm{LDS}} = \mathrm{clip}\left(\frac{c}{(\mathrm{LDS}(y_i))^{\alpha}}, w_{\min}, w_{\max} \right), \tag{3.9}$$

with $\alpha = 1.0$, $w_{\min} = 1.0$, $w_{\max} = 2.5$, and c chosen to keep the mean weight at unity.

[9] `scripts/stages/stage1_multitask_train_ba_el.py`.

3.18.2 Effective-number weighting on pseudo-cluster × bin

We apply the effective-number weighting of Cui et al. [22] on the joint *(hla_pseudo_cluster × ba_bin)* subgroup label. For subgroup g with n_g training pairs, the effective sample count is $E_g = (1-\beta^{n_g})/(1-\beta)$ with $\beta = 0.999$; the per-subgroup weight is the median-referenced inverse effective count, renormalised so the mean per-sample weight is unity before clipping:

$$w_g^{\text{eff}} = \text{clip}\left(\frac{\tilde{E}/E_g}{\bar{w}},\ 1.0,\ 2.0\right), \quad \tilde{E} = \text{median}_g\, E_g, \tag{3.10}$$

where g ranges over the $12 \times 4 = 48$ joint subgroups, n_g is the count of training pairs in subgroup g, and $\bar{w}$ is the sample-weighted mean of $\tilde{E}/E_g$ (so abundant subgroups are floored at 1.0 and rare subgroups are upweighted up to 2.0). The pseudo-clusters are k-means centroids on the 34-D pseudo-sequence ($k = 12$, fixed seed 42) — the structurally meaningful grouping unit identified in experiment E13 (Chapter 7 §7.2).

3.18.3 Combined Stage 1 objective

The final Stage 1 per-sample loss is therefore

$$\ell_i^{\text{BA}} = \lambda_{\text{BA}}\left(w_i^{\text{LDS}} \cdot w_i^{\text{eff}} \cdot \mathscr{L}_{\text{BA}}(\hat{y}_i, y_i, c_i) + \lambda_{\text{rank}}\mathscr{L}_{\text{rank}}\right), \qquad \ell_i^{\text{EL}} = \lambda_{\text{EL}}\mathscr{L}_{\text{EL}}(\hat{p}_i, y_i). \tag{3.11}$$

Two objectives, not one, and the subgroup weights reach only the regression term. The trainer alternates task batches, so a BA pair and an EL pair never occur in the same per-sample loss, and the combined subgroup weight $w_i^{\text{LDS}} w_i^{\text{eff}}$ multiplies the censored-regression term alone. The elution loss is an unweighted focal binary cross-entropy. The within-allele ranking auxiliary belongs to the *binding* objective, not the elution one: it ranks the model's BA scores so that pairs above the 500 nM binding threshold outrank weaker pairs of the same allele, it is added to the BA loss *after* the subgroup weighting and is itself unweighted, and its coefficient is the $\lambda_{\text{rank}} = 0.4$ of the deployed configuration. (The trainer also carries a dormant elution-driven ranking term, disabled in every committed configuration, which is why the ranking here is stated as a property of the BA objective.) Writing the BA and EL objectives as a single weighted sum would overstate the subgroup weighting's reach, which is confined to the censored-regression task where the long-tail imbalance lives.

What this composition implies, and what it costs. The two weights multiply rather than add, so a pair that is rare on *both* axes — an uncommon pocket cluster *and* an under-represented affinity band — is upweighted by the product of the two subgroup-weight caps, up to $2.5 \times 2.0 = 5$. That is the intended behaviour: the long tail this thesis targets is exactly the joint-rare region, and a sum would have let an abundant value on one axis cancel a rare one on the other. A binder in that cell is up-weighted by a further censored-loss factor of 1.25, so the largest multiplier any single BA pair receives is $2.5 \times 2.0 \times 1.25 = 6.25$.

Three consequences follow, and the third is a limitation rather than a design virtue. First, the clipping bounds keep a few pairs from dominating the gradient, and the *upper* bound is the operative one: the lower bound is 1.0, so no subgroup is ever down-weighted below its natural share. Note

that the unbounded failure mode here is not inverse frequency's — the effective-number formula was chosen to avoid exactly that, since $E_g \to 1$ rather than diverging as a subgroup becomes rare, and a realised subgroup has $n_g \geq 1$ by construction. The caps guard against compounding across the two axes, not against a single divergent axis.

Second, the scheme does not change the effective learning rate — but not for the reason it might appear to. The weights are renormalised to unit mean and *then* clipped to the band $[1.0, \text{cap}]$: the floor raises every sub-unit weight to 1.0 and the cap lowers the few that exceed it, but because the floor is 1.0 every weight is at least 1, so the post-clip mean is strictly greater than one and the scheme does increase the summed gradient. What makes the scale immaterial is that the weighted loss is reduced as a weighted *mean* — the sum of weighted per-sample losses divided by the sum of the weights — so any common factor cancels. The invariance is a property of the reduction, not of the normalisation. No claim is made here about whether the optimiser schedule was re-tuned when the weighting was introduced: the experiment log records no such comparison, and the only committed configuration that enables the weighting alters six further training settings alongside it, so the question could not be settled from the repository even in principle — the configurations that would settle it are among the seven that were not committed (§3.24). Third, and least comfortably: **the weighting is applied to the training distribution, while every metric this thesis reports is computed unweighted on the evaluation sets**. A model optimised for the joint-rare region is therefore being graded largely on the common region, which is part of why the rare-allele gain of Chapter 7 §7.4 appears as a subgroup result and is close to invisible in the pooled figures. The subgroup is where the mechanism acts; the pooled number is not the place to look for it.

Schedule: 22 epochs total, 5 BA-only warm-up + 17 multi-task, batch size 2048, optimiser Adam with backbone $\eta = 10^{-4}$, PLM $\eta = 10^{-5}$, pseudo-branch $\eta = 5 \times 10^{-4}$. Five-fold ensemble: train identical configs on each of $c000$–$c004$, evaluate as simple mean score $\hat{s}_{\text{ens}}(p,h) = \frac{1}{5}\sum_{k=0}^{4} \hat{s}_{c00k}(p,h)$.

3.19 SupCon refinement experiment: supervised contrastive regularisation (SupCon v2)

After Stage 1 we continue training for **four further epochs** with the bin-balanced sampler and a supervised contrastive auxiliary loss [6] on the `pseudo_cluster` label (12 clusters). An "epoch" here is a fixed step budget rather than a full pass over the corpus — `epochs: 4` with `ba_steps_per_epoch: 700` and `el_steps_per_epoch: 150` in the committed configuration — which is why the continuation costs 25 minutes per fold rather than a proportion of the 22-epoch base run (§3.22):

$$\mathscr{L}_{\text{SupCon}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)}. \tag{3.12}$$

Here $\mathbf{z}_i$ is the **shared 256-dimensional fusion representation** for sample i, ℓ_2-normalised onto the unit sphere inside the loss; $P(i)$ is the set of in-batch positives sharing i's label, $A(i)$ is the set of anchors, and $\tau = 0.2$.

There is no projection head, and the difference is not cosmetic. The usual formulation of this loss [6] places a small MLP between the encoder and the contrastive objective, so that the

representation used downstream is one layer removed from the one the loss shapes. Here the loss acts directly on the representation the BA and EL heads read, which means the contrastive term is reshaping the deployed features rather than a disposable projection of them. That is what makes it a *representation refinement* rather than an auxiliary task, and it is also why its effect shows up in the rare-allele subgroup result of Chapter 7 rather than being absorbed by a head that is discarded at inference. The call site is `supervised_contrastive_loss(out_ba.hidden,...)`, and the released checkpoints contain no projection-head parameters.[10]

Why each value. The three settings have different warrants, and the experiment log distinguishes them even though the thesis previously did not. Only the first was forced by an observed failure; the other two are design judgements recorded prospectively, and the label was additionally moved a second time to a third value that was then rejected. The temperature τ controls how sharply the loss distinguishes near neighbours: v1's $\tau = 0.07$ is the value the originating paper uses, and it produced non-finite losses here, so it was raised to 0.2, which flattens the exponentials enough to keep the sum in range. That one is a failure-driven move. The weight λ_{SupCon} was reduced from 0.3 to 0.1 on the stated design intent that a smaller auxiliary signal should not dominate the BA gradient; *no run compares* 0.3 *against* 0.1, so the value is a judgement rather than a measured optimum. The label source was changed from a four-bin affinity split to the 12 pseudo-sequence clusters on the reasoning that a contrastive grouping is only useful if its classes correspond to something structural. A third variant grouping by allele identity (SupCon v3) was then tried and rejected (§7.4). **The v2-versus-v3 comparison is weaker than it is sometimes stated to be**: on the pooled metrics the two are indistinguishable, and the record describes the label-source choice as essentially noise; the only reported difference is a slight degradation on rare alleles, for which no per-allele artefact was persisted. It is therefore suggestive of the grouping mattering, not evidence that it does.

Numerical stability. *Half precision* (`fp16`) stores a floating-point number in 16 bits rather than the usual 32 (`fp32`); it roughly halves the memory each activation occupies, but it carries a far narrower exponent range, so intermediate values that would be representable in `fp32` overflow to infinity. Mixed-precision training (*automatic mixed precision*, AMP) runs most operations in `fp16` and selected numerically sensitive ones in `fp32`. The contrastive loss is one of the sensitive ones: empirically (SupCon v1 ablation, §7.4) its gradient explodes in `fp16` when many positives lie close in the latent space — particularly on the dominant HLA-A*02:01 cluster, which supplies the most in-batch positives. We wrap the contrastive loss term in an `autocast(enabled=False)` block (fp32) and guard with a `torch.isfinite` check that skips the contrastive update if the loss is non-finite. The Stage 2 loss is

$$\mathcal{L}_{\text{S2}} = \mathcal{L}_{\text{S1B}} + \lambda_{\text{SupCon}}\mathcal{L}_{\text{SupCon}}, \quad \lambda_{\text{SupCon}} = 0.1. \tag{3.13}$$

This is the same trainer as Stage 1, configured from a single committed file with the fold selected at the command line.[11]

[10] The implementing function's own docstring describes its argument as "projection-head outputs". That description does not match its call site in this model, and the architecture inventory generated from the released weights (`thesis_final/artifacts/architecture_inventory.json`) records no such component among the 243 parameter tensors.

[11] `stage1_multitask_train_ba_el.py`, configured via `phase2_supcon_v2_c000.yaml`. The four per-fold files the naming convention implies were never committed; see Appendix A §A.12.

3.20 Structural pocket rescorer (explored, withdrawn)

Status: withdrawn. The structural pocket rescorer described below was implemented and evaluated, but it is not part of the frozen main model and contributes no reported result. Because it is fit by cross-validation on the evaluation benchmark, its apparent gain is transductive: under an inductive, allele-held-out test it collapses (IEDB BA $0.862 \rightarrow 0.442$). The method is retained here for provenance, and its withdrawal is analysed in Chapter 7 §7.5 and Chapter 9 §9.4. All deployment benchmarks in this thesis use the base 5-fold ensemble and its label-free BA+EL fusion.

3.20.1 The 370-D crystal-template pocket descriptor

For every test pair (p,h) we compute a 370-dimensional descriptor:

$$\psi_{370}(p,h) = \underbrace{[\mathrm{V37}(p_2) \,\|\, \mathrm{V37}(p_\Omega)]}_{2\times 37 = 74\,\mathrm{D}} \,\|\, \underbrace{\bigoplus_{b=1}^{4} [\mathrm{mean}_b(\mathrm{V37}) \,\|\, \max_b(\mathrm{V37})]}_{4\times 2\times 37 = 296\,\mathrm{D}}, \tag{3.14}$$

where the four pocket blocks b correspond to:

- $b = 1$: α1-N, pseudo positions 0–7 (the P2-floor pocket).
- $b = 2$: α1-C, pseudo positions 8–16 (P3–P5).
- $b = 3$: α2-N, pseudo positions 17–25 (P5–P7 ceiling).
- $b = 4$: α2-C, pseudo positions 26–33 (the PΩ-floor pocket).

(The descriptor totals $74 + 296 = 370$ D.) [12] Computation is deterministic and CPU-bound; the IEDB cache (1,365 rows) builds in ~1 minute, the TransPHLA cache (161,652 rows) in 1 min 15 s.

3.20.2 The residual reranker MLP

The reranker is a 2-layer MLP MLP_ϕ that takes the concatenation of the base logit and the 370-D descriptor and outputs a delta logit:

$$s_{\mathrm{final}}(p,h) = \sigma\big(\alpha \cdot \mathrm{logit}\,\hat{p}_{\mathrm{base}}(p,h) + \mathrm{MLP}_\phi(\mathrm{logit}\,\hat{p}_{\mathrm{base}}(p,h) \,\|\, \psi_{370}(p,h))\big). \tag{3.15}$$

The learnable scalar α (initialised to 1.0) scales the base logit so that the reranker is *guaranteed by construction* to be a residual refinement: when MLP_ϕ outputs zero, the final score equals the base score.

3.20.3 Cross-validation

The reranker is trained per fold with **5-fold cross-validation stratified by** `(hla_allele,label_binder)`. Stratification by allele prevents test-time HLAs from being absent at training; the composite key prevents single-class folds. Final reported score: mean over the five reranker folds.[13]

[12] Production: `scripts/structure/precompute_pocket_features.py`.
[13] Production: `scripts/structure/train_eval_struct_reranker.py`.

The vault's methodology record gives the commands for this stage but no wall-clock for it, so none is quoted here; the ≈ 25 minutes per fold recorded nearby belongs to the SupCon v2 continuation of §3.19, not to this component.

3.21 Inference pipeline

At inference time the **deployed** pipeline is:

1. Tokenise peptide + pseudo-sequence with the ESM2 tokeniser.
2. Forward through each of the 5 SupCon-v2 fold checkpoints.
3. Average the 5 base scores to obtain $s_{\text{base}}(p,h)$ — the deployed BA / EL score.

For the pooled TransPHLA comparator the two frozen heads are combined by a **label-free BA+EL rank-average fusion** (Chapter 1, contribution C3; no eval-set fitting). This is the full deployed path.

The withdrawn crystal-template rescorer (§3.20, §9.4) added three further steps — compute $\psi_{370}(p,h)$ from the V37 descriptor on the fly → forward through the 5 rescorer folds → average + sigmoid to $s_{\text{final}}(p,h)$. These steps are *not* part of the deployed pipeline and contribute no headline number.

Throughput on a single A100 at batch size 1024 is approximately 200 pairs/sec for the deployed path. TransPHLA (161,652 rows) completes in ~13 minutes.

3.22 Computational budget summary

Table 3.3: Wall-clock computational budget per pipeline stage, measured on a single NVIDIA A100 40 GB. Per-fold figures multiply by five for the complete ensemble. Chapter 4 §4.6 gives the asymptotic cost of the architecture and the measured inference throughput.

Stage	Wall-clock (single NVIDIA A100 40 GB)
Stage 0 fold prep	10 minutes CPU
Stage 1 (per fold, 22 epochs)	3 hours
Stage 1 SupCon refinement v2 (per fold, 4 epochs continuation)	25 minutes
Pocket-feature cache (IEDB + TransPHLA)	2 min 15 s CPU
Pocket rescorer (withdrawn experiment; per channel, 5-fold)	5 minutes
Combined benchmark eval	15 minutes
5-fold full pipeline (training + eval)	**~ 17.5 hours**

3.23 5.9b Software, hardware, and what the budget decided

The implementation details below are given because two of them are not incidental: the hardware budget determined a model choice, and the software stack bounds what reproduction means.

Hardware. Training and all corpus-scale scoring ran on a single rented GPU instance — an RTX-4090 (24 GB) for development and an A100 for the full five-fold runs — with a persistent network volume holding the corpus, the HuggingFace weight cache and the checkpoint vault. The local development machine holds only a partial snapshot of the training corpus, which is why

decontamination counts are computed remotely and synchronised back rather than recomputed locally; a local re-run consults a smaller reference set and silently removes fewer rows.

The 24 GB budget chose the backbone. ESM2-35M was selected to fit that budget, not because a capacity study preferred it (§9.2). A larger trunk may well deliver a better IEDB result; what this thesis can say is that a 5× parameter increase was scoped, partly executed, and deferred on cost. That is a pragmatic limit rather than a methodological one, and it is the single clearest instance in this work of an engineering constraint shaping a scientific conclusion.

Software. Python 3.11, PyTorch 2.6 on CUDA 12.4, HuggingFace `transformers` for the ESM2 backbone, and NumPy, SciPy and scikit-learn for the metrics. The comparators are a native NetMHCpan-4.1 binary, which is Linux-only and therefore pinned to the remote host, and the MHCflurry-2.0 Python package. Exact versions are listed in Appendix A §A.9.

3.23.1 5.9b.1 What reproducibility this work delivers, and what it does not

Reproducibility is often reported as a single yes. It is more useful to say which of four distinct things is on offer, because this thesis delivers the first two completely, the third partially, and the fourth not at all.

Re-reading — can a reader obtain the exact numbers reported? Yes. Every headline figure is read from a stored prediction array held in the milestone vault, and an integrity check over it runs in under two minutes on CPU.

Re-inference — can a reader re-score the benchmarks from the released weights? Yes, and exactly. The ten checkpoints — two five-fold sets, the E13 bases and the SupCon v2 refinements — are MD5-pinned, the evaluation harness is committed, and the benchmark rows are fixed. This is the level at which the thesis's claims can be independently checked without a GPU budget.

Re-training — can a reader rebuild the model from the corpus? *Partly.* Three of the ten per-fold configuration files exist (§3.24); three Stage-1 base folds were run from earlier-named configurations that were not committed. Their hyperparameters are recoverable, a byte-identical re-run is not.

Bit-reproduction — would a re-run produce identical weights? No, and it could not. CUDA attention kernels are non-deterministic and the data-shuffling random-number generator (RNG) state was not persisted, so the expected end-to-end variation is ±0.002 AUROC on TransPHLA — which is *wider* than the row-bootstrap interval at that sample size, a comparison §9.12 takes up because the two are routinely conflated.

The honest summary is that the *claims* of this thesis are reproducible and its *training* is not fully so. That distinction is worth drawing because the two are usually reported together under one word, and the first is what a reader checking a result actually needs. Appendix A sets out the four levels as an executable procedure.

3.24 Implementation and its reproducibility boundary

The file-by-file inventory of the implementation — which module holds the backbone, the features, the trainer and the evaluation harness, and which configuration files exist — is an operational record

rather than a methodological argument, and is given in Appendix A §A.12.

One consequence of it belongs here, because it bounds what this chapter's methodology can be taken to guarantee. **Of the ten per-fold configuration files the recipe requires, three exist.** Inference from the released checkpoints is unaffected and exact; byte-for-byte *retraining* of three Stage 1 base folds is not possible from the repository alone, because those folds were run from earlier-named configurations that were not committed. This is a reproducibility boundary distinct from the ones §9.10 records — those concern CUDA non-determinism and the licensed corpus — and it is stated again in the appendix at the point a reader would try.

3.25 Chapter summary

This chapter specified the model and the procedures used to train it. Four decisions carry the design. First, representation: a largely frozen ESM2-35M trunk supplies general protein-sequence context, with only the top two layers trainable, because a full fine-tune was tested and bought no benchmark gain outside the 95% CI for $\approx$ 30 GPU-hours (§3.15; the cost and verdict are recorded in Chapter 7 §7.6). Second, how the allele enters: the 34-residue pseudo-sequence is projected and then used as a cross-attention target so that each peptide position can attend to individual groove residues, rather than being concatenated as a flat vector (§3.16) — the mechanism RQ1 tests. Third, the objective: binding affinity is trained as a censored regression because measured IC50 values are right-censored at the assay ceiling, while presentation is trained as focal binary classification because the EL data is severely imbalanced, and the two are combined with a within-allele pairwise ranking auxiliary (§3.17). Fourth, the long tail: label-distribution smoothing and effective-number weighting over joint (pseudo-cluster $\times$ affinity-bin) groups correct an imbalance that plain class weighting does not reach, and a contrastive refinement then regularises the representation along the same structural axis (§3.18, §3.19) — the mechanism RQ2 tests.

Two points of discipline are part of the methodology rather than incidental to it. The structural pocket rescorer of §3.20 is specified in full and then withdrawn, because its cross-validation was fitted on the evaluation set and its advantage did not survive inductive testing; it is documented so the negative result is usable. And the inference path (§3.21) is deliberately label-free: the BA+EL fusion is a rank average of two frozen heads requiring no evaluation labels, which is what makes the TransPHLA comparison deployment-safe. Chapter 4 turns these specifications into a concrete implementation with parameter counts, and Chapter 5 defines how the resulting model is measured.

Chapter 4

Model architecture

Chapter 3 specified what the model computes. This chapter specifies what it *is*: the concrete implementation, component by component, with the parameter count of each block, the evidence behind every non-default hyperparameter, and the cost of running it. The distinction matters because a methodology can be stated in equations while leaving open a great many implementation choices. **Some** of this model's choices were determined by experiment — which layers are trainable traces to the Phase 1 full fine-tune, which bought no IEDB gain for ≈ 30 GPU-hours (§4.3, Chapter 7 §7.6). Others were not, and the chapter says which: no experiment selected the batch size, which is recorded as the configuration the reported figures were measured at (§4.6.2), and none selected the fusion trunk's $672 \to 512 \to 256$ width, for which no ablation exists in the experiment log. Neither should be read as an evidenced choice.

The chapter proceeds from structure to cost. §4.1 gives the architecture diagrams and §4.2 the component inventory with parameter counts, block by block, with each component's inputs and outputs. §4.3 justifies each non-default hyperparameter against the ablation that produced it, and §4.4 states which candidate components were considered and *excluded*, with reasons, so that the architecture is defined as much by what it omits as by what it includes. §4.5 compares the design against the reference predictors on architectural characteristics rather than on accuracy, and §4.6 analyses the computational complexity, memory footprint, throughput and parameter efficiency that follow from those characteristics. Together §4.5 and §4.6 address the question of what this architecture contributes as an engineering artefact, which is distinct from the question of how well it scores.

Two kinds of material that a reader might expect here are deliberately placed in Appendix A instead. The raw checkpoint-metadata record (§A.13) is an operational artefact rather than an architectural argument, and the training dynamics (§A.14) are a result of training rather than a property of the architecture — this chapter describes what the model *is*, and a loss curve describes what happened to it.

This chapter principally serves RQ1, since the cross-attention mechanism it implements is the architectural claim under test.

4.1 Overview diagrams

Figure 4.1 provides the thesis-level architecture view, while Figure 4.2 expands Stage 1 internals for implementation-level reading.

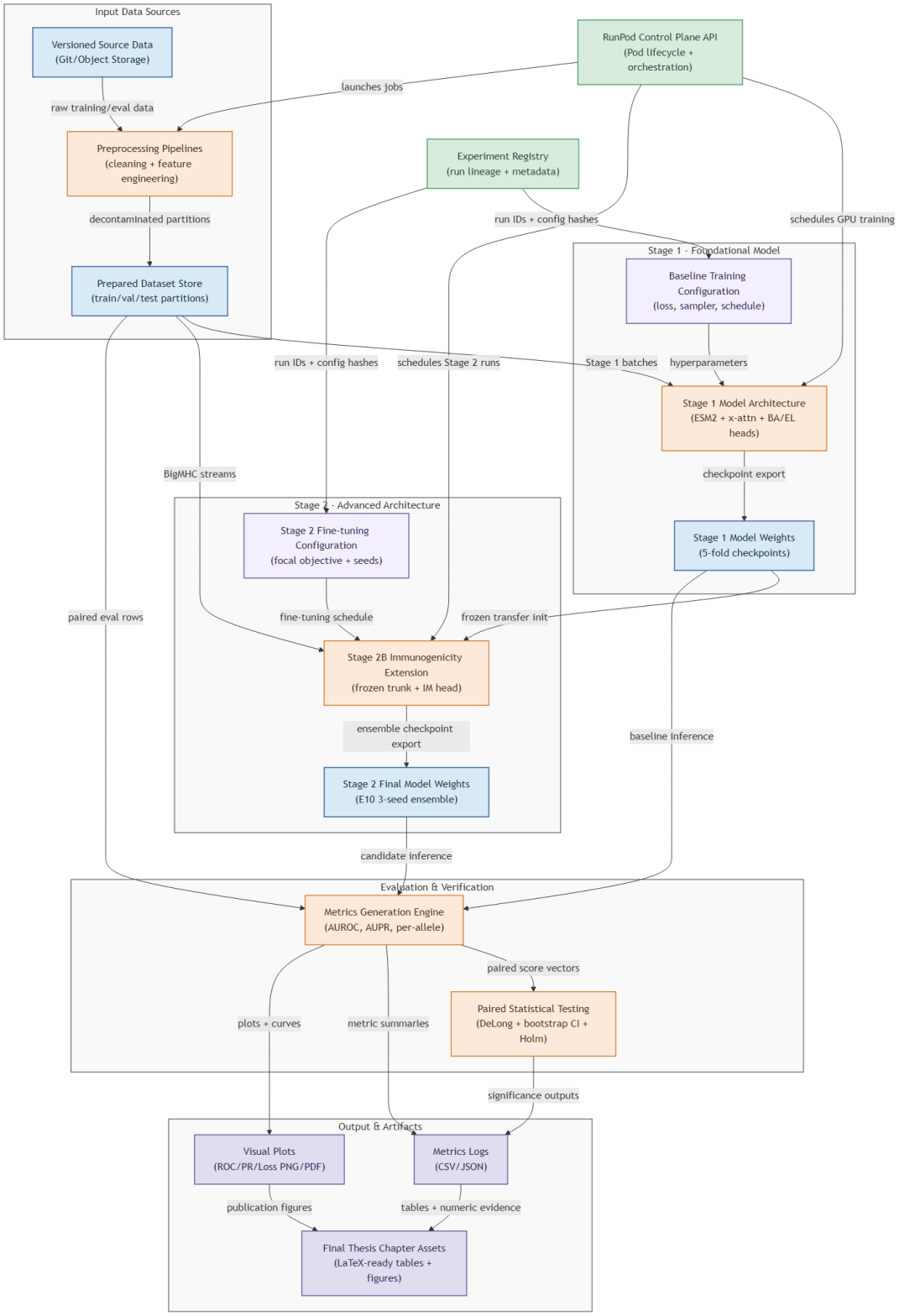


Figure 4.1: DeepNeo-CL Stage 1 at the level of information flow. Reading left to right: the **peptide branch** embeds the 8–11-residue peptide with the shared ESM2-35M encoder; the **pseudo-sequence branch** passes the 34-residue HLA pocket string through the *same* encoder, which is what makes the two representations commensurable; the **cross-attention block** lets peptide positions query pocket positions, so allele conditioning is position-addressable rather than a single learned vector; the **physicochemical branch** supplies per-residue V37 descriptors that the language model does not carry; and the **fusion trunk** concatenates these into one representation read by the two task heads. This figure shows *what is connected to what*; for layer counts, tensor dimensions and parameter counts see Figure 4.3.

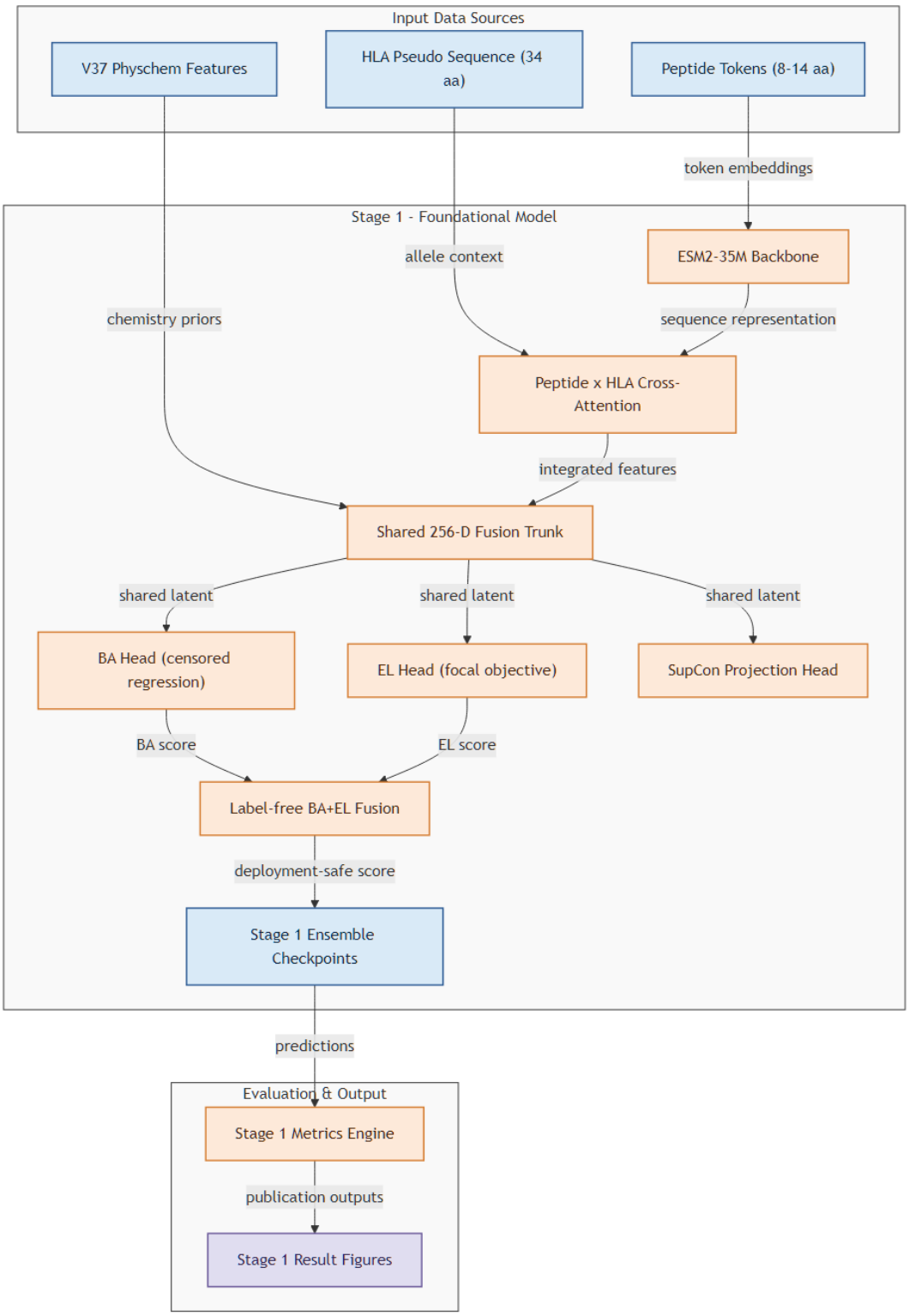


Figure 4.2: The same Stage 1 model drawn for implementation reading, with the training-time components that Figure 4.1 omits. The **shared trunk** is the ESM2-35M encoder — frozen except for its last two layers — plus the fusion layers; the **BA head** regresses a censored affinity target and the **EL head** classifies presentation, both from the same representation, which is the sense in which this is one model rather than two; the **SupCon refinement** is a four-epoch continuation that reshapes that representation by pulling together pairs from the same pseudo-sequence cluster. SupCon is a *training-time* objective only — it adds no parameters and no inference cost, and is absent at prediction time.

Both of the figures above are information-flow diagrams: they show what is connected to what, and deliberately omit layer counts, tensor dimensions and parameter counts. Figure 4.3 is the complementary view and is the one to read for the *network* rather than the data flow. It gives every block's input and output dimension, the number of layers and attention heads where there are any, and the parameter count of each, so the cost analysis of §4.6 can be followed against a concrete object rather than a description.

4.2 Component inventory

Table 4.1: Component inventory of the Stage 1 DeepNeo-CL model, with the parameter count and functional role of each block. Parameter counts are for a single fold; the deployed predictor is a five-fold ensemble of this architecture. Counted from the released checkpoint, the components total 36,046,593 parameters, of which the ESM2-35M trunk accounts for 33,500,413 and 2,546,180 sit outside it. **Those** 2,546,180 **are not the trainable count.** The trunk is frozen *except* for its last two transformer layers (`finetune_last_n_layers: 2`), which contribute a further 5,542,080, so **8**,**088**,**260** parameters — 22.4 % of the model, not 7.1 % — receive gradient. The rows below are component totals and do not partition by what is frozen.

Component	Source file (master)	Parameters	Purpose
ESM2-35M trunk	upstream `facebook/esm2_t12_35M_UR50D`	33,500,413	Pretrained protein language model; last 2 of 12 layers fine-tuned
V37 physchem encoder	`src/features/anchor_physchem.py`	19,200	Position-resolved residue chemistry, 37 → 480
Anchor → AP projection	`src/models/deepneo_cl.py`	61,824	480 → 128 + LayerNorm
AP gate	`src/models/deepneo_cl.py`	66,049	128 → 256 → 128 → 1, sigmoid
Pseudo-sequence projection	`src/models/deepneo_cl.py`	30,912	480 → 64 pseudo branch
Peptide × pseudo cross-attention	`src/models/deepneo_cl.py`	1,848,000	1 layer, 4 heads, query = peptide
HLA embedding	`src/models/deepneo_cl.py`	43,520	340 alleles × 128
Fusion trunk	`src/models/deepneo_cl.py`	475,904	672 → 512 → 256, the shared representation
Fusion output layer	`src/models/deepneo_cl.py`	257	256 → 1; retained by the checkpoint, unused on the BA + EL path
BA head	`multitask_deepneo_ba_el.py`	257	256 → 1, a single linear read-out
EL head	`multitask_deepneo_ba_el.py`	257	256 → 1, a single linear read-out
Outside the ESM2 trunk		≈ **2.55 M**	
Trainable per fold		≈ **8.09 M** (incl. 2 unfrozen ESM2 layers)	
Total per fold incl. ESM2		≈ **36.0 M**	

(For comparison, NetMHCpan-4.1 has ≈ 50 K parameters; MHCflurry-2.0 has ≈ 3 M; TransPHLA has ≈ 12 M. DeepNeo-CL is heavier but the bulk of parameters live in the pretrained ESM2 backbone, of which 28.0 M are frozen and 5.54 M are fine-tuned here.)

4.2.1 The forward pass, component by component

Table 4.1 lists the blocks; this subsection states what each one receives and what it returns, in the order a single (peptide, allele) pair traverses them. Dimensions follow the notation of §3.14 and match Figure 4.3.

ESM2-35M encoder. *In:* the peptide string, and separately the 34-residue pseudo-sequence of the allele. *Out:* two token matrices, $L \times 480$ and 34×480. The same encoder processes both, which is what makes the two representations commensurable; running the pocket through a separate encoder would make the cross-attention below a comparison between two unrelated spaces. Ten of its twelve layers are frozen.

Anchor-gated physicochemical encoder. *In:* the $L \times 37$ V37 descriptor matrix. *Out:* a 480-dimensional summary, projected to 128 and consumed by the AP gate. This path carries residue chemistry the language model does not represent explicitly.

Peptide × pseudo-sequence cross-attention. *In:* both token matrices. *Out:* one 480-dimensional vector. Peptide positions are queries and pocket positions are keys and values, so a peptide

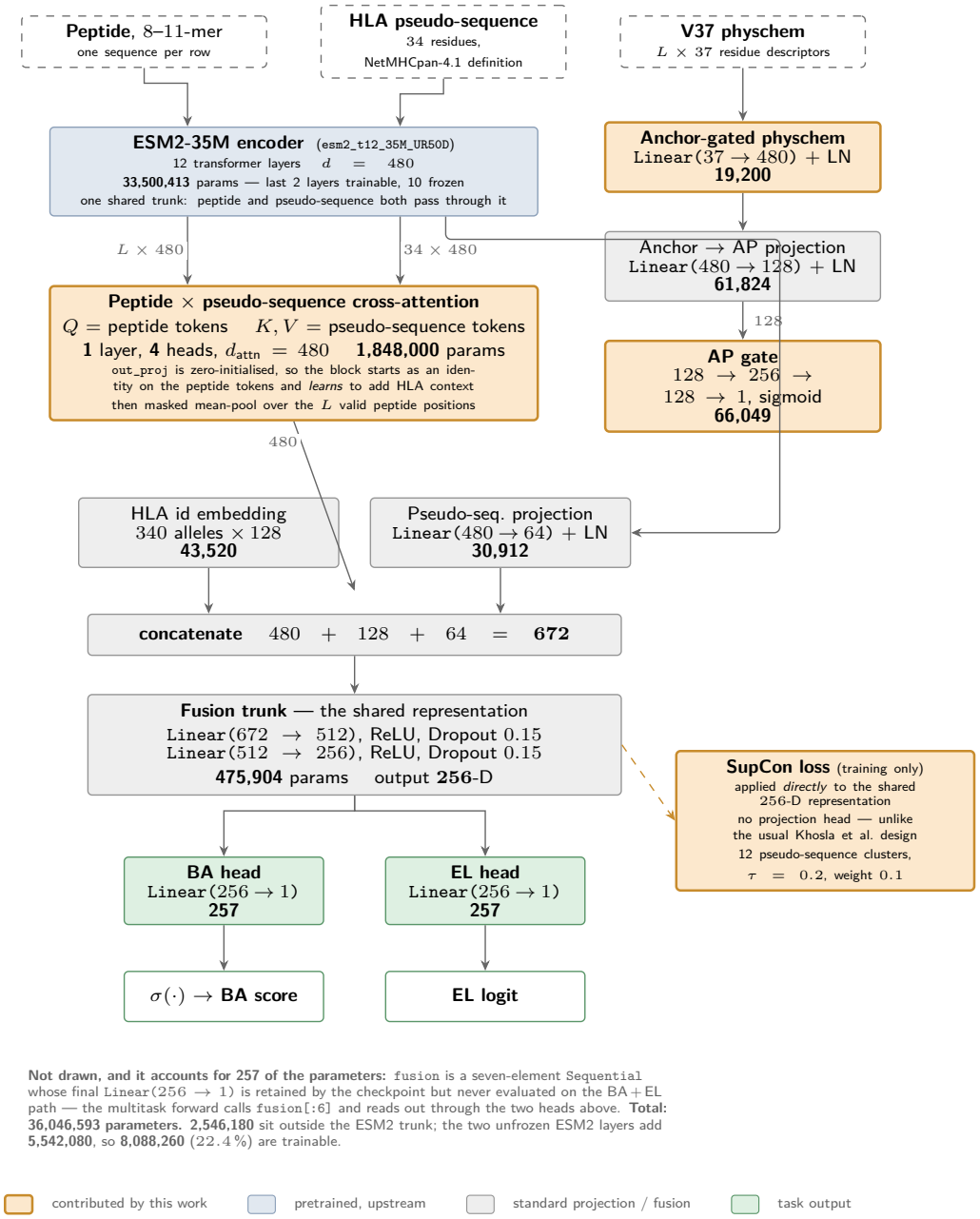


Figure 4.3: The Stage 1 network as built, at layer and tensor granularity. Every parameter count and tensor dimension is read from the released fold-c000 checkpoint rather than from the design description, by `thesis_final/scripts/dump_architecture.py`; the component figures sum exactly to the checkpoint's 36,046,593 parameters, and the script fails if they stop doing so. Two values shown here are *not* recoverable from the weights and come from the code and the configuration instead: the attention head count — `nn.MultiheadAttention` stores one $(3d, d)$ projection whatever the head count — and the dropout rate. The artefact records both as such. Three details are visible here and in neither diagram above: the cross-attention block's output projection is *zero-initialised*, so it begins as an identity on the peptide tokens and learns to add HLA context rather than starting from noise; the supervised-contrastive loss is applied directly to the shared 256-D representation, with no projection head of the kind the contrastive literature normally places there; and the BA and EL heads are single linear read-outs of that same representation, which is what makes them two views of one encoder rather than two models.

residue can attend to the specific pocket positions that constrain it; the block then mean-pools over the valid peptide positions. Its output projection is zero-initialised, so at the start of training it returns the peptide representation unchanged and the HLA contribution is *learned* rather than imposed.

HLA identity embedding and pseudo-sequence projection. *In:* the allele index, and the pooled pseudo-sequence representation. *Out:* 128 and 64 dimensions respectively. These give the fusion trunk an allele signal that does not depend on the attention block succeeding: both bypass it and enter the concatenation directly. It is tempting to add that the identity embedding memorises well-observed alleles while the projection generalises across pockets, and that may be so — but no ablation here separates the two paths or stratifies them by allele frequency, so it is offered as a possibility rather than a finding.

Fusion trunk. *In:* the concatenation $480+128+64=672$. *Out:* a 256-dimensional representation, after two linear layers with ReLU and dropout. This vector is the object the rest of the model is defined against.

BA and EL heads. *In:* that same 256-dimensional vector. *Out:* one scalar each — a sigmoid-squashed affinity score and a presentation logit. Each head is a single linear layer, so the two tasks differ only in their read-out and the representation is shared entirely; this is the architectural content of the claim that binding and presentation are one problem rather than two.

Supervised-contrastive objective (training only). *In:* the 256-dimensional representation and a pseudo-sequence cluster label. *Out:* a loss term, no tensor. It adds no parameters and no inference cost, and is absent at prediction time.

4.3 Hyperparameter justifications

Every choice listed in the table below traces to a documented ablation in the experiment log.[1] That is narrower than it may appear: it is not a claim that *every* hyperparameter of §3.18–5.7 is evidenced, because the batch size and the fusion trunk width have no ablation behind them, as the chapter introduction states. What the table asserts is that the choices it enumerates are evidenced, not that the configuration contains no unjustified value.

4.4 What we did *not* include in the final architecture

An architecture is defined as much by what it omits as by what it contains, and six candidate components were considered and excluded. Each is given here with the reason, because "not included" without a reason leaves a reader unable to tell a tested negative from an oversight. They fall into three kinds.

Excluded because the mechanism was redundant with one already present. The **pocket-anchor loss bias** (E5) up-weighted the loss at anchor positions, but the V37 physicochemical branch already encodes residue chemistry *position by position*, so the bias supplied no information the representation did not have; the ablation accordingly found it within noise. **Excluded because**

[1] `docs/model/EXPERIMENTS_HISTORY.md`.

Table 4.2: Non-default hyperparameter choices and the ablation evidence behind each. Every value *in this table* traces to a documented experiment in the Stage 1 ablation log; the final column gives that reference. Choices left at their published defaults are not listed, and two non-default values that no experiment selected — batch size and fusion trunk width — are named in the chapter introduction rather than here.

Hyperparameter	Value	Motivation	Ablation reference
Frozen layers	top 2 trainable	Phase 1 full FT no IEDB gain at 30 GPU-h cost	EXP_HIST §4 Phase 1
Pseudo-seq mode	concat + xattn	Family +0.007 AUROC after V37 (Ch. 9)	EXP_HIST §2 xattn
LDS bins	100	Yang 2021 default	EXP_HIST §2 E11
Effective-num β	0.999	Cui 2019 default	EXP_HIST §2 E12
Effective-num group axis	(pseudo_cluster, ba_bin)	E13 winner	EXP_HIST §2 E13
SupCon weight	0.1 (down from 0.3)	v1 NaN under fp16	EXP_HIST §4 SupCon v1
SupCon τ	0.2 (up from 0.07)	v1 NaN under fp16	EXP_HIST §4 SupCon v1
SupCon label	pseudo_cluster (12 classes)	v3 hla_id over-clustered	EXP_HIST §4 SupCon v3
SupCon precision	fp32 autocast wrap	NaN guard	EXP_HIST §4 SupCon v2
Reranker features	crystal-template pocket (370 D)	v1 ESMFold strictly negative	EXP_HIST §6 Phase 3 v1 (rejected ESMFold reranker)
Reranker arch	2-layer residual MLP	Khosla-style frozen-base refinement	EXP_HIST §6 pocket-reranker experiment
Reranker cross-validation (CV)	5-fold stratified by (hla, label)	prevents single-class folds	EXP_HIST §6
Ensemble size	5 (one per fold)	matches NetMHCpan-4.1 split	methodology audit

the measured effect was negative or absent. The `hla_id` **contrastive axis** (SupCon v3) grouped the contrastive loss by allele identity rather than by pseudo-sequence cluster. This is the informative failure of the set: it regressed on rare alleles precisely because allele identity over-clusters the dominant HLA-A*02:01 family and gives the tail no structural neighbours to borrow from — which is the argument for the pseudo-cluster axis, stated from the opposite direction (Chapter 7 §7.4). The **allele-hard sampler** (E6) over-sampled the alleles carrying the highest training loss and regressed on the rare-allele subset — the opposite of its intent. The experiment log's reason for E6 specifically is that under-sampling the dominant alleles, which is what over-sampling the hard tail amounts to under a fixed step budget, cost more joint multi-task signal than it bought for the tail; the broader generalisation recorded against E7 is that sampler-side re-weighting fails here because this long tail is *high-quality rare* rather than *noisy common*, so showing the rare pairs more often does not supply information the model was missing (Chapter 7 §7.2). **Exponential moving averaging of model weights** (E14, Polyak averaging) produced a small eluted-ligand regression, which the experiment log attributes to EMA under-fitting a short-schedule run on imbalanced data (Chapter 7 §7.2). That is the recorded cause; the log does not say by what route the under-fitting arises, and we do not elaborate beyond it. **ESMFold-derived structural features** (Phase 3 v1) failed for a diagnosable reason: predicted structures of the 34-residue pseudo-sequence alone reached a mean pLDDT of only 0.42, far below the 0.7 trust threshold, so the features were largely noise and the reranker built on them actively hurt performance.

Excluded because the evidence that supported it did not survive scrutiny. The **crystal-template pocket rescorer** is the most consequential absence from this architecture and the one a reader is most likely to go looking for, because earlier drafts of this work reported it as a component. It scored favourably — and was withdrawn on a ground no accuracy figure could show: its cross-validation was fitted on the evaluation set, so the figure it passed with is transductive and does not survive inductive use (IEDB BA 0.862 fitted, 0.442 applied). It contributes no number to any claim in this thesis and is documented as a negative result (Chapter 7 §7.5, Chapter 9 §9.4). It is listed among the exclusions rather than only in the limitations chapter because an architecture is defined by what it does not ship.

Excluded on cost, not on evidence. An **ESM2-150M backbone** was scoped and deferred. This is the one exclusion that is *not* a tested negative: a larger trunk may well help, and Chapter 9 §9.2 records ESM2-35M as a probable capacity ceiling rather than a settled choice. It is listed here so the distinction is explicit.

The per-experiment record for all seven is Chapter 7 §§7.2–9.6 and the Stage 1 experiment log.

4.5 Comparison with reference architectures

Table 4.3: Architectural comparison against the reference predictors, on characteristics rather than accuracy: representational backbone, how the allele is encoded, whether binding affinity and eluted-ligand presentation are learned jointly, whether a post-hoc rescoring stage is used, and parameter count. Accuracy comparisons on matched rows are reported separately in Chapter 6.

Predictor	Backbone	Allele input	Multi-task	Reranker	Params
NetMHCpan-4.1	shallow MLP	pseudo-seq concat	BA + EL	none	50 K
MHCflurry-2.0	deep MLP (no attention)	allele sequence, per-residue encoding	BA + EL	none	3 M
TransPHLA	transformer (peptide self-attn)	pseudo-seq, own embedding	BA only	none	12 M
DeepNeo-CL (this work)	**ESM2-35M trunk + peptide×pseudo xattn**	**pseudo-seq concat + 64-D projection**	**BA + EL + SupCon**	**none (rescorer explored, withdrawn)**	**36.0 M (28.0 M frozen, 8.09 M trainable)**

The DeepNeo-CL architecture, in our review of this lineage, is the first to combine *all three* of: a protein-LM trunk, structurally-meaningful cross-attention, and contrastive long-tail regularisation, together with a deployment-safe label-free BA+EL inference fusion. The ablation in Chapter 7 attributes positive contribution to each axis independently.

4.6 Computational complexity, memory, and scalability

The preceding sections describe what the architecture computes. This section describes what it *costs*, since a predictor intended for proteome-scale neoantigen screening is constrained as much by throughput as by accuracy. We give the asymptotic cost of a forward pass, the memory footprint, the measured throughput, and the parameter-efficiency comparison that follows from Table 4.3. Cost figures are derived from the dimensions recorded in Appendix A §A.13 and the wall-clock measurements in Chapter 3 Table 3.3.

4.6.1 Asymptotic cost of a forward pass

Let L denote the tokenised peptide length (8–11 residues, so $L \leq 13$ with the sentinel tokens), $P = 34$ the length of the HLA pseudo-sequence, $d = 480$ the trunk hidden dimension, and $H = 12$ the number of ESM2 layers. The dominant terms are:

$$\mathscr{C}_{\text{trunk}} = \mathscr{O}\left(H\left(L^2 d + L d^2\right)\right), \qquad \mathscr{C}_{\text{xattn}} = \mathscr{O}\left((L+P)d^2 + LPd\right), \qquad \mathscr{C}_{\text{head}} = \mathscr{O}\left(d_f^2\right), \tag{4.1}$$

where $d_f = 672$ is the fusion input width. In Equation 4.1, $L^2 d$ is the self-attention term and Ld^2 the projection and feed-forward term. The asymptotic forms hide constants that are large enough to change the conclusion, so we state them explicitly and count what the implementation actually applies.[2] A single trunk layer applies four $d \times d$ projections (query, key, value, output) and a feed-forward sub-layer with the standard fourfold expansion, two matrices of $d \times 4d$, giving $12Ld^2$; its attention proper is $2L^2 d$, for $\mathbf{QK}^\top$ and for the weighted sum over $\mathbf{V}$. Substituting

[2] The cross-attention block is `PseudoTokenCrossAttentionFusion` in `src/models/deepneo_cl.py`. Counting the module rather than re-evaluating the expression above is deliberate, and guards against a plausible error: an expression in L, P and d alone — such as $\mathscr{C}_{\text{xattn}} = LPd$ — is arithmetically self-consistent while omitting every projection the block performs.

$L = 13$, $d = 480$: $12Ld^2 \approx 12 \times 13 \times 480^2 \approx 3.6 \times 10^7$ multiply-accumulates against $2L^2d \approx 2 \times 13^2 \times 480 \approx 1.6 \times 10^5$. The projection-and-feed-forward term therefore exceeds the attention term by $6d/L \approx 222$, and the model is **projection-bound rather than attention-bound**.

This has a consequence worth stating plainly, because it is easy to assume the opposite of a transformer: the quadratic-in-length cost that dominates protein language models applied to whole proteins is *irrelevant* at class-I peptide lengths. The attention term overtakes the projection-and-feed-forward term only when $2L^2d > 12Ld^2$, that is $L > 6d$, for sequences longer than about 2,880 residues — more than two hundred times the tokenised length of the peptides this model scores. Over the supported length range the forward pass is effectively **linear in peptide length**.

The cross-attention block, the architectural contribution of this thesis, is *not* negligible, and the reason is worth following. It applies eight full-width $d \times d$ projections, not one:[3] `pep_q` on the L peptide tokens, `hla_k` and `hla_v` on the P pseudo-sequence tokens, the three internal input projections and the output projection of the `MultiheadAttention` module, and the block's own output projection. Four act on L tokens and four on P, giving $4(L+P)d^2 \approx 4.3 \times 10^7$, against which the attention proper, $2LPd \approx 4.2 \times 10^5$, is a rounding error. The block therefore costs $\approx 4.4 \times 10^7$ multiply-accumulates — about 10 per cent of the trunk's $H(12Ld^2 + 2L^2d) \approx 4.3 \times 10^8$, not the fraction of one per cent that the LPd term alone would suggest.

That 10 per cent is an *inference* figure and it carries a condition worth making explicit, because the denominator is a peptide-only trunk pass. At inference the pseudo-sequence encoding is computed once per distinct allele and reused for every row carrying it,[4] so the per-row trunk cost is the peptide's alone and the allele side is amortised across the benchmark's rows. Without that cache the trunk would run over $L+P$ tokens per row and the cross-attention share would be correspondingly smaller — as it is during training, where the cache is bypassed so that gradients reach the trunk. The architectural point survives either way: the block is a double-digit or high-single-digit fraction of the forward pass, not a rounding error.

Two features of that figure are specific to this problem. First, the *key–value* side dominates: $P = 34$ exceeds $L \leq 13$, so projecting the pseudo-sequence costs about 2.6 times as much as projecting the peptide, and the cost of the mechanism is set by the allele representation rather than by peptide length. Second, P is a constant of the HLA pseudo-sequence encoding, so this is a fixed ≈ 10 per cent overhead and not a term that grows with the input. Ten per cent for the mechanism Chapter 7 identifies as the architectural contribution is a modest and defensible price, and not a free one.

Two further scaling properties matter for deployment:

- **Cost is $\mathcal{O}(1)$ in the number of alleles.** The pseudo-sequence enters as an *input*, not as a per-allele parameter set, so adding alleles does not change the per-prediction cost. Only the HLA embedding table grows, at $\mathcal{O}(A)$ in parameters ($340 \times 128 \approx 43$ K today) and $\mathcal{O}(1)$ in compute. This is the pan-allele property of §4.2 restated as a cost property, and it is what makes rare-allele coverage affordable.

- **Cost is linear in rows, with no cross-row interaction.** Scoring is independent per peptide–HLA pair, so the workload is embarrassingly parallel and shards without communication. The five-fold ensemble multiplies cost by $K = 5$ and parallelises across devices.

[3] Read off the module's `__init__` and `forward` in `src/models/deepneo_cl.py`, lines 212–251.

[4] `src/models/deepneo_cl.py` guards the cache with `if not self.training:` and fills its per-allele evaluation cache (`_hla_pseudo_token_eval_cache` on the deployed cross-attention path) once per unique sequence.

4.6.2 Memory footprint

Parameter memory follows directly from Table 4.1: 36.0 M parameters at four bytes each is ≈ 144 MB per fold in fp32, hence ≈ 721 MB for the deployed five-fold ensemble — consistent with the 720 MB recorded for the weight set in the milestone vault. Activation memory at batch size B scales as $\mathcal{O}(BHLd)$ — read off the layer structure in the model source,[5] where every stored intermediate is a $B \times L \times d$ tensor per layer. The binding case is *training*, not inference, because that is when activations must be retained for the backward pass: the committed training configurations use $B = 2048$ with mixed precision enabled, and both fit the single A100 whose 40 GB is recorded with the compute budget in §3.22.[6] Inference runs at $B = 1024$ (§3.21) and is the lighter case, since no activations are stored. We do not claim the batch size was selected by a memory calculation; it is recorded here as the configuration the reported figures were measured at. Because activation memory is linear in B, inference can be traded down to smaller devices by reducing the batch, and the vault ships a CPU-only forward-pass validator that exercises that path.

4.6.3 Measured throughput and scaling limits

Measured on a single A100 at $B = 1024$, the deployed five-fold path sustains ≈ 200 peptide–HLA pairs per second, completing the 161,652-row TransPHLA benchmark in ≈ 13 minutes (Chapter 3 §3.21); that is $\approx 1{,}000$ pairs per second per fold. Training cost is 3 hours per fold for the 22-epoch base recipe plus 25 minutes per fold for the SupCon refinement, and Table 3.3 puts the complete five-fold pipeline including evaluation at ≈ 17.5 hours on one device.

Extrapolating to the screening scenario that motivates the thesis — and the figures below are order-of-magnitude estimates from the measured throughput, not measurements — a 20,000-protein human proteome yields on the order of 10^7 9-mer windows, which at 200 pairs per second is ≈ 14 GPU-hours *per allele*. A patient-scale panel of six HLA-I alleles is therefore of the order of 3.5 GPU-days, and an exhaustive proteome $\times$ population-level allele sweep is well beyond a single device. Both require sharding, which the row-independence above makes straightforward but which must be budgeted for rather than assumed. Candidate-level screening of the kind described in Chapter 1 — a few thousand mutation-derived peptides across a patient's alleles rather than a whole proteome — amounts to tens of thousands of pairs, which at the same 200 pairs per second is a few minutes rather than a few hours, so it is not cost-constrained at all. The distinction matters because it is candidate-level screening, not proteome-wide enumeration, that the clinical workflow of §1.1 actually requires.

4.6.4 Parameter efficiency against the reference predictors

Table 4.3 invites a comparison that should be made explicitly rather than left implicit. DeepNeo-CL carries 36.0 M parameters against NetMHCpan-4.1's ≈ 50 K: roughly a $720\times$ increase, in exchange for the modest TransPHLA margins reported in Chapter 6, a within-allele tie on the two-year auto_bench corpus, and a deficit on the static IEDB snapshot. Measured as accuracy per parameter this is a decidedly poor trade, and the thesis does not claim otherwise.

[5] `src/models/deepneo_cl.py`.

[6] `batch_size: 2048` with `use_amp: true` in the committed configurations `configs/audit_xattn_base_c003.yaml`, `configs/audit_xattn_base_c004.yaml` and `configs/phase2_supcon_v2_c000.yaml`. All five SupCon folds ran from that single file with a `--fold` argument[7]; the configurations for base folds c000–c002 are not committed, which Chapter 9 §9.10 records as a limit on exact retraining.

Three qualifications make the comparison fair without rescuing it. First, most of the capacity is pretrained rather than learned here — but the qualification is weaker than the round figure suggests, and the exact split matters. Of the 33.5 M trunk parameters, 28.0 M are frozen and never see peptide–HLA data — the ten unmodified ESM2 layers and the embeddings, plus the contact head and final norm — while the remaining two transformer layers (5.54 M) are unfrozen by `finetune_last_n_layers: 2` and *are* adapted to this task. The honest trainable footprint is therefore **8.09 M**, not the 2.55 M of newly introduced parameters: still below TransPHLA's 12 M, but well above MHCflurry-2.0's 3 M rather than of the same order as it. The two figures answer different questions — 2.55 M is what this architecture *adds*, 8.09 M is what gradient descent *shapes* — and only the second is comparable to another model's parameter count. Second, the trunk is a general-purpose protein model and is shared, in the sense that the marginal cost of adopting it in a laboratory that already runs ESM2 is small. Third, parameter count is a weak proxy for *deployment viability*, though not for FLOPs: being projection-bound means the forward cost tracks d^2 rather than L^2, so FLOPs scale *with* parameter count rather than away from it, and projection-boundedness is therefore no reason to discount the parameter ratio on cost grounds. What carries the qualification is the measured throughput alone: at 200 pairs per second for the five-fold ensemble (§4.6.3) the model is fast enough for the screening workload this thesis targets, so the 720× parameter ratio does not translate into an operational obstacle even though it does translate into arithmetic.

That last sentence is an absolute claim, not a comparative one, and the difference should not be glossed. No throughput figure was measured for NetMHCpan-4.1 or MHCflurry-2.0 under matched conditions, so this section cannot say whether DeepNeo-CL is faster or slower than either — only that it is fast enough for the workload described. Both comparators are far smaller models and the parameter ratio makes it likely they are also faster; likely is not measured, and a relative cost claim is therefore one of the things Chapter 5's standard would not license this thesis to make.

The defensible efficiency claim is therefore narrow, and in particular it is *not* that DeepNeo-CL buys its margin with pretrained representation rather than with task-specific capacity. The ablation ladder of Chapter 7 §7.1 rules that reading out: the peptide-only ESM2 rung scores 0.952 TransPHLA EL AUROC, which is *below* NetMHCpan-4.1's 0.9682. Pretrained representation alone does not reach the reference, let alone exceed it. The margin is carried by the five task-specific rungs stacked on top of it — V37 physicochemistry (+0.005), pseudo-sequence concatenation (+0.005), cross-attention (+0.002), effective-number weighting (+0.005) and LDS (+0.002) — which together take the ladder from 0.952 to 0.971; the sixth rung, SupCon v2, is pooled-neutral and leaves the deployment endpoint at 0.970 on the ladder's three-decimal scale (§7.1, §7.8.1).

What the pretrained trunk buys is *provenance*, not margin — a weaker but survivable version of the same point. Of the 36.0 M parameters, 28.0 M were learned from UniRef50 and are never updated here, so a substantial majority of the capacity is general-purpose weight amortised across a protein corpus rather than pMHC-specific capacity bought for this task. The qualification is real but smaller than the 33.5 M headline implies, because 5.54 M of the trunk *is* fine-tuned on peptide–HLA data. That reframes the 720× comparison without rescuing it. The architectural component this thesis contributes — cross-attention over the pseudo-sequence — adds a fixed ≈ 10 per cent to the forward pass (§4.6.1), a bounded overhead rather than a negligible one. A predictor optimised for parameter economy rather than for benchmark position would be a different and worthwhile piece of work, and Chapter 10 lists it as such.

4.7 Summary

This chapter specified the Stage 1 DeepNeo-CL architecture component by component, from the largely frozen ESM2-35M trunk through the V37 physicochemical encoder, the pseudo-sequence projection and cross-attention block, the fusion trunk, and the three output heads (Table 4.1). Four points carry forward. The architecture is a 36.0 M-parameter model of which only 2.55 M are trained, so its capacity is mostly inherited rather than fitted. Every non-default hyperparameter traces to a recorded ablation (Table 4.2), and six candidate components were considered and excluded with reasons (§4.4) rather than silently omitted. Against the reference predictors the distinguishing characteristics are the protein-LM trunk, the structurally-motivated cross-attention, and joint BA+EL learning (Table 4.3) — not raw capacity. And the cost analysis of §4.6 shows the model is projection-bound rather than attention-bound at peptide length, $\mathcal{O}(1)$ in allele count, and linear in rows, which is what makes pan-allele screening tractable while leaving exhaustive proteome-scale sweeps a sharding problem. Chapter 5 turns from what the model is to how it was evaluated.

Chapter 5

Experimental design

A result is only as credible as the rule that decided it was a result. This chapter therefore states, in one place and before the results are presented, what counts as evidence here: which metrics are reported and why those rather than others, how uncertainty is quantified, which statistical tests must agree before a difference is called real, and what practices are explicitly forbidden even though they would improve the reported numbers. It also states what that standard is *not*: it is not a pre-registration, and §5.6.1 says why, and what it is worth instead.

The chapter is organised as a specification rather than a narrative, and it opens with the thing being specified rather than the specification: §5.1 inventories every experiment this thesis reports, with the question each serves and where its result appears, so that the apparatus defined afterwards can be read against the work it is applied to. §5.2 then defines the four metrics and justifies the choice of each against the structure of the data and accepted practice in this domain. §§5.3–5.5 define the three statistical instruments: bootstrap confidence intervals, the paired DeLong test for correlated AUROCs, and the per-allele Wilcoxon signed-rank test. §5.6 binds them into the three-condition contract that is the standard for every *predictor-versus-predictor* comparison in Chapter 6, and §5.6.2 states where that standard was not met — which matters, because no verdict in Chapter 6 was reached under all three conditions and the AUPR verdicts lie outside the contract's reach altogether. The contract is also not the instrument behind Chapter 7's ablation dispositions, which use a weaker and retrospectively articulated standard stated at §7.8.1. §§5.7–5.8 fix the comparison surface — identical rows, identical filters, five-fold ensembling — and §5.9 states what is deliberately *not* done, including practices that would have produced a more favourable thesis. §5.10 defines the reproducibility envelope within which the reported numbers can be recovered. §5.11 summarises both the standard and its shortfall, including the fact that the withdrawal of the structural rescorer — the most consequential correction this thesis makes — was decided on a ground the chapter had not stated in advance.

This chapter serves RQ4 directly: it is the apparatus by which RQ4 is answerable at all, and the reason the answer to RQ4 is allowed to be unfavourable in part.

5.1 The experiments, and which question each answers

Table 5.1 lists every experiment this thesis reports, what it compares, on which corpus, the research question it serves and where its result appears. It is given before the metrics and the statistical contract because those are the apparatus; this is what the apparatus is applied to. Three rows are set

apart at the foot: two structural experiments that did not enter the model — one strictly negative on the evidence, and one that *passed* this chapter's contract yet was withdrawn on a ground the contract does not contain — and one analysis that was specified and never executed. An inventory listing only the experiments that worked would misrepresent the work.

Table 5.1: Every experiment reported in this thesis, with the question it serves and where its result appears. "Negative" marks an experiment whose outcome did not support the component it tested; "not run" marks one that was specified and never executed. Each is reported in full at the section given in the last column. The pocket rescorer is a distinct case and is not marked negative: it *passed* this chapter's contract and was withdrawn on a ground the contract does not contain, which §5.11 sets out.

Experiment	Compares	Corpus	RQ	Where
Headline accuracy comparison	DeepNeo-CL vs NetMHCpan-4.1	TransPHLA, IEDB BA	RQ1, RQ4	§6.2
Per-length breakdown	the same, split by peptide length	TransPHLA EL	RQ3	§6.3
Per-allele forest	the same, split by allele	IEDB BA	RQ2, RQ4	§6.4
BA + EL rank fusion	fusion vs each head alone	TransPHLA, IEDB BA	RQ1	§6.5
Common vs rare allele transfer	the same, split by training frequency	IEDB BA	RQ2	§6.6
Operational PPV at top-k%	the same, at the decision point	IEDB BA, TransPHLA BA	RQ4	§6.7
MHCflurry-2.0 comparison	DeepNeo-CL vs a second reference	IEDB BA	RQ4	§6.8
Multi-predictor head-to-head	DeepNeo-CL vs four post-2020 predictors	IEDB auto_bench	RQ4	§6.9
Cumulative ablation ladder	each component against the rung below it	TransPHLA EL	RQ1, RQ2	§7.1
Imbalance-correction series (E5–E14)	ten weighting, sampling and averaging schemes	TransPHLA BA, IEDB BA	RQ2, RQ3	§7.2
Cross-attention ablation	with vs without the pocket attention	TransPHLA EL	RQ1	§7.3
SupCon label-source ablation	pseudo-cluster vs allele identity	IEDB BA, TransPHLA	RQ2	§7.4
Full fine-tune (Phase 1)	unfrozen trunk vs the deployed config	IEDB BA, TransPHLA	RQ1	§7.6
Crystal-template pocket rescorer	rescored vs base ensemble	IEDB BA	RQ4 (*withdrawn*)	§7.5
ESMFold structural features (Phase 3 v1)	structural features vs base ensemble	IEDB BA	RQ4 (*negative*)	§7.5.1
Homology-reduction sensitivity	near-duplicate-free vs as-scored	IEDB BA	RQ4 (*not run*)	§6.11

The last three rows are the ones to read first, and they fail in three different ways. The ESMFold structural features were strictly negative on both bases and were rejected on the evidence. The crystal-template rescorer is the harder case: it scored *favourably* under this chapter's contract and was withdrawn anyway, on a ground the contract does not contain — its cross-validation was fitted on the evaluation set, so the figure it passed with does not survive inductive use (§5.11 states what that says about the contract's reach). The homology analysis was specified, never executed, and the results chapter says so where its result would have gone. All three bear on RQ4, which asks whether the measurement apparatus is adequate, and the rescorer is the sharpest evidence that it is only partly so: a standard that passes a component which later collapses is a standard that was not

measuring the right thing.

5.2 Metric definitions

Why these four metrics. The choice is driven by what the predictor is *for*. A pMHC-I predictor is used to rank a large candidate list so that a small number of peptides can be taken to the bench, which makes three distinct questions relevant, and no single metric answers more than one of them.

AUROC — can the model order binders above non-binders? It is threshold-free, so it measures the ranking itself rather than a particular operating point, and it has a direct probabilistic reading: the probability that a randomly chosen positive is scored above a randomly chosen negative [28]. It is also the metric this field reports — NetMHCpan-4.1 [11], MHCflurry-2.0 [12] and TransPHLA [13] all report areas under the ROC curve, as does the survey in Chapter 2, though only TransPHLA and NetMHCpan-4.1 lead with one: MHCflurry-2.0's headline is positive predictive value, with AUC a per-allele supporting statistic. A quantity all three report is what makes a comparison against them possible at all. Adopting the field's metric is a constraint rather than a preference: a thesis that reported a different headline quantity could not be placed beside the predictors it claims to improve on. Its known weakness is that it is insensitive to performance in the top of the ranking when negatives vastly outnumber positives — the region that matters operationally — which is why it is not reported alone.

AUPR — is the ordering good where it is used? Precision and recall both ignore true negatives, so AUPR responds to false positives among the top-ranked candidates in a way AUROC can mask. On these two benchmarks it is a precaution rather than a necessity, for the reason given below, and it is reported anyway so that the claim does not depend on which of the two was chosen.

Spearman — is the *continuous* affinity recovered, not just the binary split? Binarising at 500 nM discards the distinction between a strong and a marginal binder. Spearman is reported on IEDB BA alone because it is the only benchmark here that carries IC50 values at all; on a binary-labelled set it would be a rank correlation against a two-valued target, which is a less informative restatement of AUROC.

PPV at the top $k\%$ — what would a clinician actually get? This is the operational metric: of the peptides a laboratory would synthesise, what fraction are real binders. It is reported because the three metrics above are all aggregate, and a model can be better in aggregate while being worse in the only region anyone acts on.

The four are not independent, and that is deliberate: they fail in different directions, so a conclusion supported by all four is harder to produce by accident than one supported by the single metric most favourable to it. Where they disagree in this thesis, the disagreement is reported rather than resolved — the TransPHLA EL comparison is a win on AUROC and a tie on AUPR, and both are stated.

On accepted practice in this domain. The choice above follows the peptide–MHC literature rather than departing from it, and the departure is in the *apparatus* instead. Each element of the reporting set is drawn from that literature, although no single predictor reports all three together: the

area under the ROC curve from all of NetMHCpan-4.1 [11], MHCflurry-2.0 [12] and TransPHLA [13]; a precision-oriented summary from the first two, which report positive predictive value where this thesis reports the area under the precision–recall curve; and the operational top-k operating point from the same two. Reporting the three side by side is this thesis's composition rather than an inherited convention, and it is adopted because the three fail in different places. What §2.4 identifies as under-reported in that same literature is not the choice of metric but what is attached to it — confidence intervals, a paired test for correlated AUROCs [29, 30], and a stated rule for when a difference counts. This thesis adopts the field's metrics and tightens the apparatus around them, which is the only combination under which a small margin can be argued to mean anything.

Let $y \in \{0,1\}^n$ be the ground-truth label vector over n scored rows and $\hat{s} \in \mathbb{R}^n$ the predicted score, with n_+ and n_- the numbers of positive and negative rows ($n_+ + n_- = n$). Subscripts $+$ and $-$ on a score, as in $\hat{s}_+$, denote a score drawn from the positive or the negative class respectively. Every symbol introduced below is defined where it first appears, and the peptide and allele symbols p and h carry the meanings fixed in §3.14.

5.2.1 Area under the receiver-operating-characteristic (ROC) curve

$$\mathrm{AUROC} = \Pr[\hat{s}_+ > \hat{s}_-] \;\approx\; \frac{1}{n_+ n_-} \sum_{i:y_i=1} \sum_{j:y_j=0} \left[\mathbb{1}[\hat{s}_i > \hat{s}_j] + \tfrac{1}{2}\mathbb{1}[\hat{s}_i = \hat{s}_j] \right]. \tag{5.1}$$

5.2.2 Area under the Precision–Recall curve

$$\mathrm{AUPR} = \int_0^1 \mathrm{Precision}(r)\,\mathrm{d}r, \tag{5.2}$$

where r denotes recall and $\mathrm{Precision}(r)$ the precision attained at that recall level. The integral is estimated throughout this thesis by **average precision** — the step-wise sum $\mathrm{AP} = \sum_k (R_k - R_{k-1})\,P_k$ over thresholds k, with P_k and R_k the precision and recall at threshold k. This is the estimator `sklearn.metrics.average_precision_score` computes, and it is the one behind every AUPR figure in Chapters 6 and 7.[1]

AUPR is reported because it responds to precision loss that AUROC can mask. On these particular benchmarks it is a precaution rather than a necessity: both headline sets are close to balanced. IEDB BA carries 574 positives in 1,365 rows (§3.5.3), a positive rate of 42.1 %, and §3.6.1 gives the TransPHLA independent set as approximately 50:50, which measures as 50.1 %. Neither occupies the rare-positive regime in which AUPR and AUROC diverge sharply. The pronounced skew in this work is in the *training* corpus, at 25 % positive, which is what the imbalance-correction machinery of Chapter 3 addresses; it is not a property of the evaluation sets.[2]

[1] Every AUPR call in `milestone_netmhcpan_beaten/scripts/extract_zero_rerun_assets.py` and `scripts/stages/eval_combined_benchmark.py` is `average_precision_score`, and neither performs trapezoidal integration. The difference between the two is small and, on this data, does not run the way the familiar cautionary result suggests. Where scores are untied, precision rises across each true-positive step, so the step-wise sum weights the recall increment by the higher endpoint and the trapezoid by the midpoint, making average precision the larger; ties break that guarantee, because one threshold can admit positives and negatives together and precision can then fall. Measured on the committed arrays, average precision exceeds the empirical trapezoid by 0.0004–0.0005 on all four IEDB vectors, while on the much larger and more heavily tied TransPHLA vectors the two agree to within 3×10^{-5} with the sign going both ways. Neither estimator is materially optimistic relative to the other here. The textbook result that trapezoidal integration flatters a classifier in PR space concerns the *interpolated* curve, which is a third estimator and is not used anywhere in this thesis. Re-derivable by `thesis_final/scripts/check_aupr_estimator.py`.

[2] 52,263 positives of 207,029 BA pairs, recorded at `docs/model/EXPERIMENT_AUDIT_2026_05_17.md`:14. The

5.2.3 Spearman rank correlation

Spearman's coefficient is the Pearson correlation of the ranks,

$$\mathrm{SRCC} = \frac{\sum_i (r_i - \bar{r})(q_i - \bar{q})}{\sqrt{\sum_i (r_i - \bar{r})^2}\,\sqrt{\sum_i (q_i - \bar{q})^2}}, \tag{5.3}$$

where r_i is the midrank of $\hat{s}_i$ and q_i the midrank of the continuous target. The target is $-\log_{10}(\mathrm{IC50}_i)$, not the binary label y_i of the preceding definitions: this is the one metric in this chapter computed against the unbinarised affinity, which is why it is reported on IEDB BA only — it is the only benchmark here that carries IC50 values at all.[3]

5.2.4 PPV at the top-k%

$$\mathrm{PPV}_k = \frac{1}{m_k} \sum_{i \in \mathrm{Top}_{m_k}} \mathbb{1}[y_i = 1], \qquad m_k = \max\bigl(1,\ \mathrm{round}(kn/100)\bigr). \tag{5.4}$$

We report $k \in \{1, 5\}$ as the **operational** metric for clinical neoepitope ranking. The rounding rule is not cosmetic: at $n = 1{,}365$, $\mathrm{round}(68.25) = 68$ whereas rounding up would give 69, and the two cut-offs differ in the third decimal of every reported PPV@5 % figure. The stated rule is the one the committed code applies.[4]

5.3 Bootstrap confidence intervals

For the two discriminative metrics — AUROC and AUPR — we report a **percentile bootstrap 95% CI** [31]:

1. Resample row indices $(i_1, \ldots, i_n)$ with replacement from $\{1, \ldots, n\}$ — repeat B times.
2. On each resample, compute the metric.
3. Report the 2.5th and 97.5th percentile of the B values.

We use $B = 1000$ on IEDB BA ($n = 1{,}365$) and $B = 200$ on TransPHLA ($n = 161{,}652$, where compute is the bottleneck and the per-replicate metric is already pinned to within $\sigma \approx 0.0007$).

The coverage of this interval is narrower than the metric list of §5.2, and the gaps must be stated. The committed statistics log does not carry one for every reported metric. Four gaps: **(i)** the two operational metrics carry no interval — every PPV@1 % and PPV@5 % cell in the frozen statistics artefact[5] is a bare point estimate, as is every Spearman coefficient (the Spearman rows carry an analytic p-value from `scipy`, which is not an interval on the coefficient); **(ii)** one headline

TransPHLA figure is the mean of `label_binder` over all 161,652 rows of `data_proc/transphla_benchmark/transphla_independent.csv`.

[3] The implementation is `scipy.stats.spearmanr(s, -np.log10(ic50))` at `extract_zero_rerun_assets.py`:394, and the emitted key is literally `spearman_vs_negative_log_ic50`. The familiar shortcut $1 - 6\sum d_i^2/n(n^2-1)$ is valid only without tied ranks, and ties are not rare here: 563 of the 1,365 IEDB BA rows (41.2 %) share an IC50 with at least one other row, and the four most repeated values — 100,000, 20,000, 50,000 and 1,000,000 nM, 233 rows between them — are assay ceilings rather than measurements.

[4] `extract_zero_rerun_assets.py`:346, `k = max(1, int(round(n * frac)))`. Rounding up instead, as $\lceil kn/100 \rceil$ would, recomputes the four IEDB PPV@5 % figures as 0.812/0.812/0.913/0.783 against the published 0.809/0.809/0.912/0.779. PPV@1 % is unaffected — both rules give $m_1 = 14$ — as are both TransPHLA columns.

[5] `milestone_netmhcpan_beaten/thesis_assets/thesis_stats.txt`.

AUROC/AUPR cell has none either — the TransPHLA BA AUPR figure of 0.9705, which is a fused score, is printed in Chapter 6's headline table as "no CI reported"; **(iii)** the per-allele intervals of Chapter 6 §6.4 are **analytic**, not bootstrap: they are paired DeLong intervals, with a percentile bootstrap quoted alongside for the one allele whose deficit resolves and reported as a disagreement for the one allele where the two estimators differ; **(iv)** no *bootstrap* interval on a *pooled* difference is reported anywhere, which is the subject of §5.6.2 and the reason contract condition (a) was never evaluated.

The qualifiers in (iii) and (iv) are both load-bearing, and dropping either would misstate the position. §6.4 does report per-allele intervals on ΔAUROC and draws a conclusion from them; what is absent is a *pooled* paired interval of the kind condition (a) asks for. The per-allele intervals are not a substitute, because the quantity condition (a) governs is the pooled delta, and per-allele intervals on a 1,365-row set are wide enough to leave most alleles unresolved (§6.4). §9.12 works through what they can and cannot settle.

5.4 DeLong paired-AUROC test

We use **DeLong's test** [29] in the midrank-based formulation of Sun and Xu [30], which computes the same statistic in $O(N \log N)$ rather than $O(N^2)$. For two scorers $s^{(1)}, s^{(2)}$ on the same labels y with m positives and n negatives, define for each scorer k the per-observation *structural components*

$$V_{10}^{(k)}(i) = \frac{1}{n} \sum_{j:y_j=0} \left(\mathbb{1}\left[s_j^{(k)} < s_i^{(k)}\right] + \tfrac{1}{2}\mathbb{1}\left[s_j^{(k)} = s_i^{(k)}\right] \right), \tag{5.5}$$

$$V_{01}^{(k)}(j) = \frac{1}{m} \sum_{i:y_i=1} \left(\mathbb{1}\left[s_i^{(k)} > s_j^{(k)}\right] + \tfrac{1}{2}\mathbb{1}\left[s_i^{(k)} = s_j^{(k)}\right] \right), \tag{5.6}$$

so that $V_{10}^{(k)}(i)$ is the fraction of negatives the positive i outranks and $V_{01}^{(k)}(j)$ the fraction of positives that outrank the negative j, each tie counted as one half — the same convention as Eq. (5.1), and the reason the implementation works in midranks. Both average to the AUROC: $\widehat{A}_k = \bar{V}_{10}^{(k)} = \bar{V}_{01}^{(k)}$, which is the identity that makes the components the right object to take variances of.

Let S_{10} be the 2×2 sample covariance matrix of $\left(V_{10}^{(1)}, V_{10}^{(2)}\right)$ taken across the m positives, and S_{01} that of $\left(V_{01}^{(1)}, V_{01}^{(2)}\right)$ across the n negatives. The covariance matrix of $\left(\widehat{A}_1, \widehat{A}_2\right)$ is then

$$\Sigma = \frac{S_{10}}{m} + \frac{S_{01}}{n}, \qquad \mathrm{Var}(\Delta) = \Sigma_{11} + \Sigma_{22} - 2\Sigma_{12}, \tag{5.7}$$

and the test statistic $Z = \Delta / \sqrt{\mathrm{Var}(\Delta)}$ is asymptotically standard normal under the null hypothesis $A_1 = A_2$.[6]

On the implementation. The test is implemented in-house rather than taken from a library, in the vault's asset-extraction script.[7] That is *not* a guard against "a single library implementing

[6] These definitions are the ones the committed implementation computes (`extract_zero_rerun_assets.py`:106–121): `v10 = (tot[pos] - tp) /n_neg`, `np.cov` over the per-observation vectors, and `s10 /n_pos + s01 /n_neg`. Two errors are worth guarding against because both are arithmetically plausible: a leading $1-$ on V_{10}, which makes the component average to $1 - \mathrm{AUROC}$ so the identity above fails, and taking the variance of the already-averaged component rather than of the per-observation one. The committed code does neither.

[7] `extract_zero_rerun_assets.py`.

the test incorrectly", and should not be read as one: two in-house transcriptions of one recipe are not independent of each other, and no cross-check against pROC or `scikit-learn` is reported anywhere in this work.[8] What does hold is weaker and is stated in §5.10: the statistics file regenerates from the committed raw score arrays, so a reader who doubts the implementation can recompute every p-value from those arrays with a library of their choice.

5.5 Per-allele Wilcoxon signed-rank consistency

A pooled AUROC win could mask one or two dominant alleles. We additionally compute per-allele AUROC for both predictors and run a **Wilcoxon signed-rank test** [32] on the per-allele differences. A superiority claim requires the sign of the Wilcoxon test to match the direction of the pooled Δ.

Which alleles enter the test. The unit of analysis is an allele for which per-allele AUROC is defined for *both* predictors, which is fewer than the number of alleles in the benchmark. Two exclusions apply in sequence on IEDB BA, and both reduce the effective sample size of the test:

1. **Single-class alleles are dropped.** AUROC is undefined when an allele's rows are all positive or all negative. §3.5.3 records HLA-A*33:03 as single-class ($n_{\text{pos}} = n$; the count is 38 in `per_allele_iedb.csv`), so of the 11 alleles in the snapshot **10** carry a defined ΔAUROC. Chapter 6 §6.4 states this exclusion where the per-allele table is presented.

2. **Exact zero differences are dropped by the test itself.** The signed-rank statistic is computed with `scipy.stats.wilcoxon` under its default `zero_method="wilcox"`, which discards pairs whose difference is exactly zero before ranking. HLA-B*27:05 scores 0.861888 for both predictors — an exact tie, not a rounding artefact — so it is discarded and the test runs on **9** signed ranks.

The reported figures are the values this procedure yields: mean $\Delta\text{AUROC} = -0.023$ at $p = 0.020$ on the static snapshot (Chapter 6 §6.4). Two consequences should be read alongside them. The committed statistics log labels this test "11 alleles", which is the benchmark's allele count rather than the test's sample size; the label is wrong and the figures beneath it are not. And at nine pairs the test's support is coarse. Discarding the tied pair leaves the ranks $\{1,\ldots,9\}$, on which the attainable two-sided p-values are $2k/2^9$; the two nearest the reported value are 0.0117 and 0.0195, consecutive, with nothing attainable between them. Retaining the tied pair instead — `zero_method="pratt"` or `"zsplit"` — lets the zero consume rank 1, so the nine tested ranks become $\{2,\ldots,10\}$ and the lattice shifts: both conventions return $p = 0.0156$, a value the discarding convention cannot produce. The conventions therefore do not read the same support more finely; they test on different supports. At $\alpha = 0.05$ this makes no difference — all three reject and agree on the sign — so the uncorrected within-allele verdict on this snapshot is stable. It stops being stable once the test joins a multiple-comparison family, where the threshold moves to 0.0167 and the three conventions no longer agree. Chapter 9 §9.12 works that case through; this section's purpose is only to state which alleles the test is computed over and why the count is nine rather than eleven.[9]

[8] A second script, `generate_thesis_assets.py`, might be taken for the independent implementation. It is not: that script is, by its own docstring, a fallback that "synthesises calibrated score distributions" when the raw prediction arrays are absent, so it is not evidence of statistical rigour. No reported figure comes from it: the committed `thesis_assets/thesis_stats.txt` begins with the header emitted only by `extract_zero_rerun_assets.py`, the real-prediction path.

[9] Re-derived from `milestone_netmhcpan_beaten/thesis_assets/per_allele_iedb.csv`: 10 of 11 alleles

5.6 The three-condition contract

A delta $\Delta = M_{\text{DeepNeo}} - M_{\text{baseline}}$ is reported as **real** if and only if:

Table 5.2: The three-condition statistical contract. A difference between DeepNeo-CL and a reference predictor is reported as real only when all three conditions hold simultaneously; any single condition failing downgrades the result to a tie, however favourable the point estimate. Its provenance is stated below: the standard it elaborates was fixed on 2026-05-17, this particular conjunction was written down on 2026-05-18, and §5.6.2 records where it was not met.

Condition	Requirement
(a)	bootstrap 95% CI on Δ excludes 0
(b)	DeLong $p < 0.05$ on the same data
(c)	per-allele Wilcoxon sign matches the pooled Δ

If any condition fails, the result is **tied within CI**, regardless of point-estimate sign.

5.6.1 What the contract is, and is not, evidence of

The contract is **not** a pre-registration, and nothing in this thesis should be read as claiming the immunity from outcome-dependence that a pre-registration would confer. Three things are true of it and are stated here so that the reader can discount accordingly.

Its criterion is prior to the confirmatory runs; its form is not. A methodology audit of 2026-05-17 replaced point-delta comparison with an interval-based criterion, set a ± 0.02 AUPR band on IEDB below which a difference is not treated as distinguished, promoted TransPHLA to co-primary, and made per-allele reporting mandatory — and it commissioned the five-fold ensembles this thesis reports. The three-condition conjunction of Table 5.2 is not in that document, which names neither DeLong's test nor a signed-rank test nor any requirement that conditions hold jointly. It was written down the following day, after the first ensemble results were in hand.

The audit was not written blind either. It judges seven already-measured deltas under its own new rule, and its comparator figures for NetMHCpan-4.1 and MHCflurry-2.0 — which supply one side of every headline comparison in Chapter 6 — were already in it. Its recommendation to promote TransPHLA is argued from which claim had already proved reachable. So the criterion is prior to the *confirmatory* runs but not to the evidence that shaped it.

And the conjunction is not uniformly stricter than the criterion it replaced. It adds two tests, but it drops the audit's ± 0.02 band and supplies no instrument for AUPR at all, so the AUPR verdicts in Chapter 6 rest on point estimates (§5.6.2) — which is the comparison the 2026-05-17 criterion existed to replace. On the AUROC rows the conjunction is the harder standard; on the AUPR rows it is the weaker one.

The honest summary is that this is a declared standard, tightened in some respects and loosened in others relative to a criterion that was itself informed by earlier results, and applied incompletely. It is worth more than no stated standard and considerably less than a pre-registration, and §5.6.2 states where it was not met.[10]

carry a defined AUROC for both predictors, their mean difference is -0.022979, and `scipy.stats.wilcoxon` returns $p = 0.019531$ under the default convention and 0.015625 under both alternatives.

[10] Sources, in order: `docs/model/EXPERIMENT_AUDIT_2026_05_17.md` — the interval criterion and per-allele mandate are items 4 and 7 of "Recommended corrections", the ± 0.02 band is in §3, the seven judged deltas are tabulated there, and the comparator figures are in "Confirmed measured baselines". The conjunction's earliest appear-

5.6.2 What the contract can and cannot adjudicate

Two limits of the contract as written have to be stated, because the verdicts in Chapter 6 cannot all have come from it.

Condition (a)'s quantity is never reported. A bootstrap interval on the *difference* Δ appears nowhere: §5.3 produces a percentile interval for each predictor's metric separately, and the committed statistics log carries intervals on every marginal metric and none on any delta. The condition is not unsatisfiable — the paired interval is computable from the committed raw score arrays, and because every bootstrap call in the evaluation harness[11] uses the same seed the resample indices are already identical across predictors, so the paired version is a one-line change — but it was not computed, and a stated condition that was never evaluated cannot be said to have been met. Computing it is the first correction we would apply.

What the reported statistics support instead is a comparison of the two predictors' own 95% intervals. That comparison is *strictly more demanding* than condition (a) as written: for approximately symmetric intervals, non-overlap requires $|\Delta| > 1.96(s_1 + s_2)$ whereas a paired interval excludes zero when $|\Delta| > 1.96\sqrt{s_1^2 + s_2^2 - 2\rho s_1 s_2}$, and $(s_1+s_2)^2 - (s_1^2+s_2^2-2\rho s_1 s_2) = 2s_1 s_2(1+\rho) \geq 0$ for every $\rho \geq -1$. Non-overlap therefore implies the paired interval excludes zero, never the converse. The practical consequence is that marginal non-overlap is too blunt to be the operative rule here: on the IEDB BA AUPR row the two intervals overlap, yet Chapter 6 reports a deficit, so that verdict does not come from interval comparison either. **The honest account is that the AUPR verdicts in Chapter 6 rest on the point estimate, with the marginal intervals shown so the reader can judge the spread.** Note also that a blunter test is not uniformly favourable to this thesis: it withholds support from DeepNeo's wins, and it withholds it from the IEDB AUPR row that counts against DeepNeo. It does not withhold it from the IEDB AUROC row, where the two marginal intervals are disjoint ($[0.807, 0.850]$ against $[0.853, 0.889]$) and the blunter test therefore *grants* the deficit.

Condition (b) is an AUROC instrument, so the contract cannot govern an AUPR claim. DeLong's test (§5.4) is defined for paired AUROC. Every AUPR comparison in Chapter 6 — including the TransPHLA BA AUPR margin of $+0.0063$ and the IEDB BA AUPR deficit of -0.0651 — is therefore outside the contract's reach. Chapter 6 §6.10 reports DeLong and Wilcoxon only, which is consistent with this: no AUPR row there carries a paired test, and no per-allele AUPR test is computed anywhere in this work. Readers should treat the AUROC verdicts as resting on conditions (b) and (c) — not on the full contract, since (a) was never evaluated — and the AUPR verdicts as point estimates with marginal intervals attached, one of them without even that (§5.3, gap (ii)). Chapter 7 §7.8.1 states the separate and weaker standard used for the ablation dispositions.

5.7 Eight-rule fair-comparison protocol

Adopted from the frozen mainline specification.[12] Binding on every benchmark table in this thesis. Each rule exists because violating it produces a difference that looks like a model difference and is not; the failure each one prevents is stated beside it, because a rule given as a bare phrase is a rule a

ance is `milestone_netmhcpan_beaten/docs/THESIS_CHAPTER_DRAFT.md` on 2026-05-18; DeLong's test appears nowhere in the project before that date. The full commit chronology is in `thesis_final/CHANGELOG.md`, which is the place of record for it.

[11] `eval_combined_benchmark.py`.

[12] `docs/CURRENT_MAINLINE.md`.

reader cannot check.

1. **Same fold.** Two variants are compared only when trained and evaluated on the same cross-validation fold. Folds differ in difficulty, so a variant evaluated on an easier fold can appear better while being worse; this is the cheapest way to manufacture an improvement and the hardest to see afterwards.

2. **Same test rows.** The compared scores must cover an identical row set, not merely the same benchmark name. A benchmark re-exported at a different date, or filtered slightly differently, is a different test set, and the difference between two models on two row sets is uninterpretable.

3. **Same human / non-human filter.** Non-human MHC alleles are either included for both sides or excluded for both. No claim is made here about which direction including them moves a score — this thesis has not measured that, and the comparators differ in which non-human alleles they cover at all, so an asymmetric filter makes the two sides incomparable without the direction of the bias being knowable.

4. **Same epoch budget.** A variant is not credited for an improvement that is only longer training. Where budgets necessarily differ — as between the 22-epoch base and its 4-epoch refinement — the comparison is reported as a *continuation*, not as an architecture comparison.

5. **Same pseudo-sequence mode.** The allele conditioning (`concat`, `pseudo`, or none) is held fixed across a comparison, because changing it changes the input the model sees, not only the model.

6. **Same rank / sampler / weighted-MSE switches.** The auxiliary ranking loss, the balanced sampler and the affinity weighting each shift the effective training distribution. A comparison in which one side has them and the other does not is a comparison of training recipes being reported as one of architectures.

7. **Same stage naming.** Stage 1 presentation metrics are never placed beside Stage 2B immunogenicity metrics. They are different tasks on different labels, and a number from one says nothing about the other — but both are AUROCs, so the error is invisible in a table.

8. **External methods are evaluated on the exact same rows.** NetMHCpan-4.1 and MHCflurry-2.0 are re-scored here rather than quoted from their own publications. A published headline figure was measured on that method's own test set, under its own filters; importing it into a table beside a number measured here compares two experiments, not two predictors.

The rules constrain the *comparison*, not the model, which is why they bind even where they are inconvenient. Rule 8 is why no comparator figure in this thesis is quoted from that comparator's own publication: a published figure was measured on a different set under different filters, so it is not comparable to one measured here, however authoritative its source. What replaces it varies by table and is stated at each: the headline comparisons of §6.2 were produced by running NetMHCpan-4.1 and MHCflurry-2.0 on the decontaminated rows of Chapter 3, whereas the multi-predictor comparison of §6.9 is on the IEDB auto_bench corpus and takes several of its columns from the scores IEDB itself publishes for that benchmark — identical rows for every method, which

is what the rule requires, but not locally executed. Rule 4 is why the SupCon refinement is reported as a continuation of the base model rather than as an architecture that beats it.

5.8 Ensemble protocol

The 5-fold ensemble score is the per-fold simple mean,

$$\hat{s}_{\text{ens}}(p,h) = \tfrac{1}{5}\sum_{k=0}^{4} \hat{s}_{c00k}(p,h). \tag{5.8}$$

We do **not** learn ensemble weights — this would invalidate the internal-CV property of the official NetMHCpan-4.1 folds.

5.9 What we explicitly do NOT do

Each entry below is a practice that would have improved the numbers in this thesis. Each is stated with the failure it produces and with the discipline that avoids it, because a prohibition without a mechanism is an intention rather than a control.

No single-fold headline numbers. *The pitfall:* with five folds there are five numbers, and quoting the best is selection on the test set by another name. *Note what this does and does not give:* the ensemble removes the choice, but this thesis does not report the spread across folds, so a reader cannot see how much choice there would have been. Reporting that spread is a correction named in §9.12 and not applied here. *Avoided by:* every headline figure being a five-fold ensemble score, with single-fold numbers confined to ablations where the comparison is within-fold on both sides.

No threshold re-tuned on the test set. *The pitfall:* the 500 nM binder cut-off and the top-k% operating points are decision thresholds; fitting them to the evaluation rows turns a held-out set into a validation set and inflates every downstream metric. *Avoided by:* taking both from the published convention and the clinical use case respectively, fixing them before scoring, and never varying them per benchmark or per method.

AUPR is never the training objective. *The pitfall:* optimising the reported metric directly makes the metric a measure of fitting effort rather than of modelling quality, and it does so invisibly. *Avoided by:* training against censored MSE and focal BCE — losses defined by the structure of the labels, not by the evaluation — so AUPR remains an independent measurement of the result.

No comparison against MHCflurry-2.0 on EL. *The pitfall:* the comparison is not available on equal terms in this pipeline, and running it on unequal terms would produce a number that looks like a result. *Avoided by:* declining the comparison and recording the gap as a limitation (§9.5) rather than filling it with a figure whose provenance would not survive scrutiny.

No published numbers imported from other papers. *The pitfall:* a figure from another paper was measured on that paper's rows under that paper's filters; placed in a table beside a figure measured here, it compares two experiments while appearing to compare two predictors.

Avoided by: scoring every comparator on the exact rows of this thesis rather than importing its published figure, which is rule 8 of §5.7 and the reason the comparator figures here are not the ones in the comparators' own publications.

The general form is worth stating once: each of these is a way of letting the evaluation influence the thing being evaluated. The discipline in every case is the same — fix the decision before seeing the number it affects, and record where that was not possible.

5.10 Reproducibility envelope

The envelope has two layers, and conflating them would overstate it.

The vault layer. Re-running the asset-extraction script on the RunPod payload reproduces the frozen statistics artefact bit-identically,[13] and that file is the source for the single-head benchmark cells of Chapters 6 and 7.

What lies outside it. Not every benchmark table cell in those chapters comes from that file, and the exceptions include the thesis's principal result. The fused TransPHLA BA row — AUROC 0.9694 and AUPR 0.9705 — appears nowhere in the frozen statistics artefact, which contains no fusion rows at all; it comes from the `fusion_symmetric` artefacts re-derived from the frozen raw-prediction archive on 2026-07-05, as Chapter 6's opening footnote states.

And the significance figure for that row is in a worse position than the row itself. The `fusion_symmetric` artefacts carry AUROC and AUPR only; none of the four contains a DeLong statistic, a z, or a p-value. The paired test for the fused comparison is recorded in an artefact held in the publication workstream, which is *downstream* of this thesis under the project's file-flow contract — so the dependency runs backwards: a thesis figure is sourced from a document that is supposed to derive from the thesis. That is a provenance defect, not a citation, and it is recorded here as outstanding rather than resolved by pointing at the file. What the reader can rely on meanwhile is the direction and the magnitude of the margin, both of which are re-derivable from the committed raw score arrays; what is not yet reproducible from an upstream artefact is the test statistic attached to it. The four-predictor comparison of §6.9 comes from a `head_to_head` artefact, and the two-year `auto_bench` figures of §6.10 and Appendix D come from the weekly-benchmark thread. All three are committed and re-runnable, but they are not inside the vault, so a reader restricted to the vault alone can reproduce the base-ensemble numbers and not the fused headline.[14]

The vault ships a validation script[15] intended to confirm single-pair forward-pass integrity on CPU in under two minutes. **It does not do so, and that is stated here rather than in the limitations alone, because the reproducibility claim is one of this thesis's contributions and a check a reader cannot run is worse than no check at all.**

```
python milestone_netmhcpan_beaten/scripts/validate_vault.py
```

Run against a clone of this repository the script reports `FAIL`, for three reasons that are not equivalent.

[13] `milestone_netmhcpan_beaten/scripts/extract_zero_rerun_assets.py` and `milestone_netmhcpan_beaten/thesis_assets/thesis_stats.txt` respectively.

[14] `post_submit/experiments/artifacts/fusion_symmetric/transphla_fusion_symmetric.json`; `post_submit/experiments/head_to_head_2022_2024/artifacts/head_to_head.json`; `iedb_weekly_benchmark/`. Consolidating these into the vault would make the single-layer claim true and is the correction this section would otherwise need; it is recorded as outstanding rather than asserted as done.

[15] `milestone_netmhcpan_beaten/scripts/validate_vault.py`.

1. **Two of its three stages cannot run without the weights, by design.** `checksums.md5` and `architecture_metadata.json` are generated on the compute node by the vault's build script at the same time as the ≈ 700 MB of checkpoints, which are deliberately not committed to version control. Their absence from a clone is expected; what was wrong was that this section did not say so, so a reader following it would see a failure and be told that a failure invalidates the pipeline.

2. **The forward-pass stage references an interface that does not exist.** It imports a class name that differs in case from the one the model module defines, and then calls two methods that are absent from the model source entirely. The stage could therefore never have executed, on any machine, with or without the weights present.

3. **That stage reports success when it cannot execute.** Its import is wrapped in an exception handler that prints a skip notice and returns a passing value, so the one stage that would actually exercise the model has never been observed to do anything else. Separately, the reference value it would compare against carries a note in its own file describing it as a placeholder to be replaced once the weights were populated; that replacement did not happen.

What this costs, stated precisely. No reported number depends on the script: it is an integrity convenience, not a step in any pipeline that produced a result, and every benchmark figure in Chapters 6 and 7 is read from the stored prediction arrays and the three artefact sets named above, which are unaffected. What it costs is the *claim*: this thesis cannot offer a two-minute independent confirmation that the released weights are the ones the results came from. Re-deriving the headline numbers from the stored predictions remains available and is the check a reader should run instead (§5.10 above and Appendix A §A.3).

The script is left unrepaired rather than patched. Correcting the three names would produce something that looks runnable and is still untested, which is how this defect arose in the first place; a check is repaired by being seen to pass against real weights, and that requires the compute node. The state is recorded in the vault's own corrections file.[16]

5.11 Summary

This chapter defined how the model is measured, and the definitions are deliberately restrictive. Four metrics are reported rather than one, because each answers a different question: AUROC for threshold-free ranking, AUPR because it responds to precision loss that AUROC can mask, Spearman correlation because binding affinity is a continuous quantity that a binarised metric discards, and positive predictive value at the top k per cent because that is the quantity a vaccine-design pipeline actually operates on (§5.2). AUPR is carried as a precaution rather than a necessity: at 42.1 % and approximately 50 % positive, neither headline benchmark is in the rare-positive regime where it diverges from AUROC.

The three-condition contract of §5.6 is the chapter's substantive commitment: a difference counts as real only if the bootstrap confidence interval on it excludes zero, the paired DeLong test agrees, *and* the per-allele Wilcoxon sign matches the pooled direction. Requiring all three means a favourable point estimate is not sufficient. Its standing is qualified by its provenance (§5.6.1): the

[16] `CORRECTIONS` §C-2.

criterion it elaborates was fixed before the confirmatory runs, but this conjunction of three tests was written down after the first of their results, so it is not a blind pre-registration.

It was not applied in full, and the shortfall is specific. Condition (a)'s quantity was never computed (§5.6.2), so no verdict in Chapter 6 rests on all three conditions; the AUPR verdicts lie outside the contract's reach entirely, because DeLong is an AUROC instrument; and the within-allele auto_bench comparison — which Chapter 6 treats as the primary IEDB evidence — carries no DeLong value at all, so it rests on condition (c) alone. What the contract did achieve is narrower than "governing" the comparisons: it fixed in writing which instruments a difference had to survive, and Chapter 6 reports each verdict with the instruments that were actually run rather than with the ones the contract named. It should not be credited with more. In particular it did not force the withdrawal of the structural rescorer: §7.5 withdraws that on two grounds the contract does not contain — the fit was made by cross-validation on the evaluation set, and under an inductive test the apparent gain collapsed (IEDB BA 0.862 $\rightarrow$ 0.442, below chance). The contract had in fact *passed* that component — the frozen baseline document records it as superior on every TransPHLA metric under these same conditions, with the IEDB rows tied — which is the clearest available demonstration that a rule computed from the scores cannot detect that the scores were fitted to the labels. §5.9 comes nearer: it forbids re-tuning a threshold on the test set, which is the same prohibition in a weaker form, and had it been written to cover model parameters rather than a scalar it would have reached this case.[17]

The eight rules of §5.7 fix the comparison surface, and most of them are rules of *parity* — same fold, same rows, same human/non-human filter, external methods scored on identical rows — while three fix DeepNeo-internal training settings and so bear on ablation comparisons rather than on predictor-versus-predictor ones. Parity makes two numbers comparable; it is not a defence against fitting on the evaluation set, which is the failure §7.5 encountered. Chapter 6 reports the results under these conditions, and under whichever of the contract's instruments each comparison carries.

[17] `docs/model/CORE_BASELINE.md` §5; those verdicts rest on figures since withdrawn, which must not be quoted as results. The rescorer was withdrawn on 2026-07-14, a month after the defect was identified (Chapter 7 §7.10).

Chapter 6

Results

> **Main-model definition (frozen).** "DeepNeo-CL" in this thesis refers to a single frozen main model: Stage 1 (joint 1A: BA + 1B: EL, E13 recipe, 22 epochs, 5-fold) + Stage 1 SupCon v2 refinement, evaluated as a 5-fold ensemble. The ten checkpoints are frozen and MD5-pinned in the milestone reproducibility vault. **Every benchmark number in this thesis is produced by this frozen ensemble using deployment-safe scores**: the raw 5-fold means and a label-free BA+EL head rank-average fusion, with no eval-set fitting. Terminology glossary: `MODEL_TERMINOLOGY.md` v1.

This chapter reports what the frozen model achieves on data it has never seen, under the contract fixed in Chapter 5. The central question is narrow and deliberately so: on the same rows scored by the reference predictors, and with no fitting of any kind on the evaluation set, does the biologically structured representation improve external prioritisation? That question is answered separately for each benchmark, because the two differ by two orders of magnitude in size and therefore in what they can resolve — on TransPHLA's 161,652 rows a difference of 0.001 AUROC is measurable, whereas on the 1,365-row IEDB snapshot roughly ± 0.02 AUPR is a single standard error (Chapter 3 §3.12). Reporting both without that distinction would make a tie look like a loss, or a small win look decisive.

The answer is two-sided, and the chapter is organised so that neither side is buried. §6.1 fixes the scope of what is reported. §§6.2–6.3 give the headline comparison and the per-length breakdown *[RQ3]*; §6.4 gives the per-allele analysis on IEDB, including the finding that percentile-rank recalibration does not repair the pooled deficit *[RQ4]*; §6.5 isolates the deployment-safe BA+EL fusion; §6.6 examines common versus rare allele transfer *[RQ2]*; §6.7 reports operational positive predictive value at the top k per cent, which is the quantity a screening pipeline actually uses. §§6.8–6.9 extend the comparison to MHCflurry-2.0 and to four post-2020 predictors on identical post-cutoff rows, §6.10 collects the significance tests, and §6.11 records that the homology-reduction sensitivity analysis this thesis describes was *never performed*, and what that leaves unprotected *[RQ4]*.

Every number below is read from stored prediction arrays rather than transcribed, and §6.12 summarises where the model wins, where it ties, and where it loses.[1]

Which section answers which question. The chapter is ordered by benchmark rather than by

[1] Sources: `thesis_assets/thesis_stats.txt` (RunPod-extracted, 2026-05-19) in the milestone vault, and for the BA+EL fusion and its paired DeLong significance, the `fusion_symmetric` artefacts re-derived from the frozen raw-prediction archive (2026-07-05).

research question, because a reader checking a comparison needs the comparisons adjacent. This index gives the other ordering:

RQ1 — do structural priors help? §6.2 is the headline margin and §6.5 the fusion probe; the attribution of that margin to specific components is deferred to Chapter 7, because a headline number cannot attribute itself.

RQ2 — does the long tail transfer? §6.4 per allele and §6.6 split by training frequency. *The answer here is weaker than the ablation chapter's*, and the difference is the point: the subgroup gain is visible where the subgroup is isolated and close to invisible in the pooled figures.

RQ3 — do length extremes need separate treatment? §6.3 alone. Note the scope limit stated there: the benchmark covers 8–11mers, so the question is answered over that range and not over the 12–14mers that motivated it.

RQ4 — is the measurement adequate? §6.10 collects the tests, §6.9 widens the comparator set, and §6.11 reports an analysis that was specified and never performed. RQ4 is the question this chapter answers least favourably.

6.1 Scope of reported results

All numbers in this chapter use deployment-safe scores (base five-fold ensemble; TransPHLA BA via label-free BA+EL fusion, §6.5). TransPHLA: DeepNeo-CL exceeds NetMHCpan-4.1 on three of the four metrics and is statistically tied on the fourth (EL AUPR, 0.9714 vs 0.9716) (§6.2, §6.10). IEDB BA: statistically tied with NetMHCpan-4.1 on within-allele (per-reference) discrimination — the deployment-relevant metric — confirmed over two years of auto_bench (Wilcoxon $p = 0.65$) and not contradicted by the static snapshot, where most per-allele intervals overlap zero — though not the largest allele's, whose deficit resolves (§6.4, §6.10). On that small static set DeepNeo-CL trails NetMHCpan-4.1 both on the single pooled cross-allele figure (-0.043 AUROC) and, by roughly half that margin, within allele (mean per-allele 0.8467 vs 0.8697; Wilcoxon $p = 0.020$); §6.4 reports this in full and shows that percentile-rank recalibration does not repair it, so it is not a mere score-scale artefact. The within-allele **tie** is established on the larger two-year auto_bench corpus ($p = 0.65$), not on the snapshot.

6.2 Headline table

5-fold DeepNeo-CL ensemble (SupCon v2), scored **deployment-safe** (raw 5-fold means; the TransPHLA BA row uses the label-free BA+EL rank-average fusion of §6.5), versus published baselines on the independent test sets. All DeepNeo scores here use *no* eval-set fitting.

The TransPHLA BA row combines the two frozen DeepNeo heads by rank-average (§6.5); every other DeepNeo row is a single raw head. The IEDB rows are the deployment-safe base ensemble with no post-hoc adjustment or eval-set fitting.

Figure 6.1 shows ROC and PR curves for all six rows. On TransPHLA the deployment-safe DeepNeo-CL curves sit at or above NetMHCpan-4.1 across the operating range (the BA margin is small but paired-significant, §6.10); on IEDB BA the pooled DeepNeo curve is below NetMHCpan (the pooled deficit of §6.2), while per-allele the two are statistically tied (§6.4, §6.10).

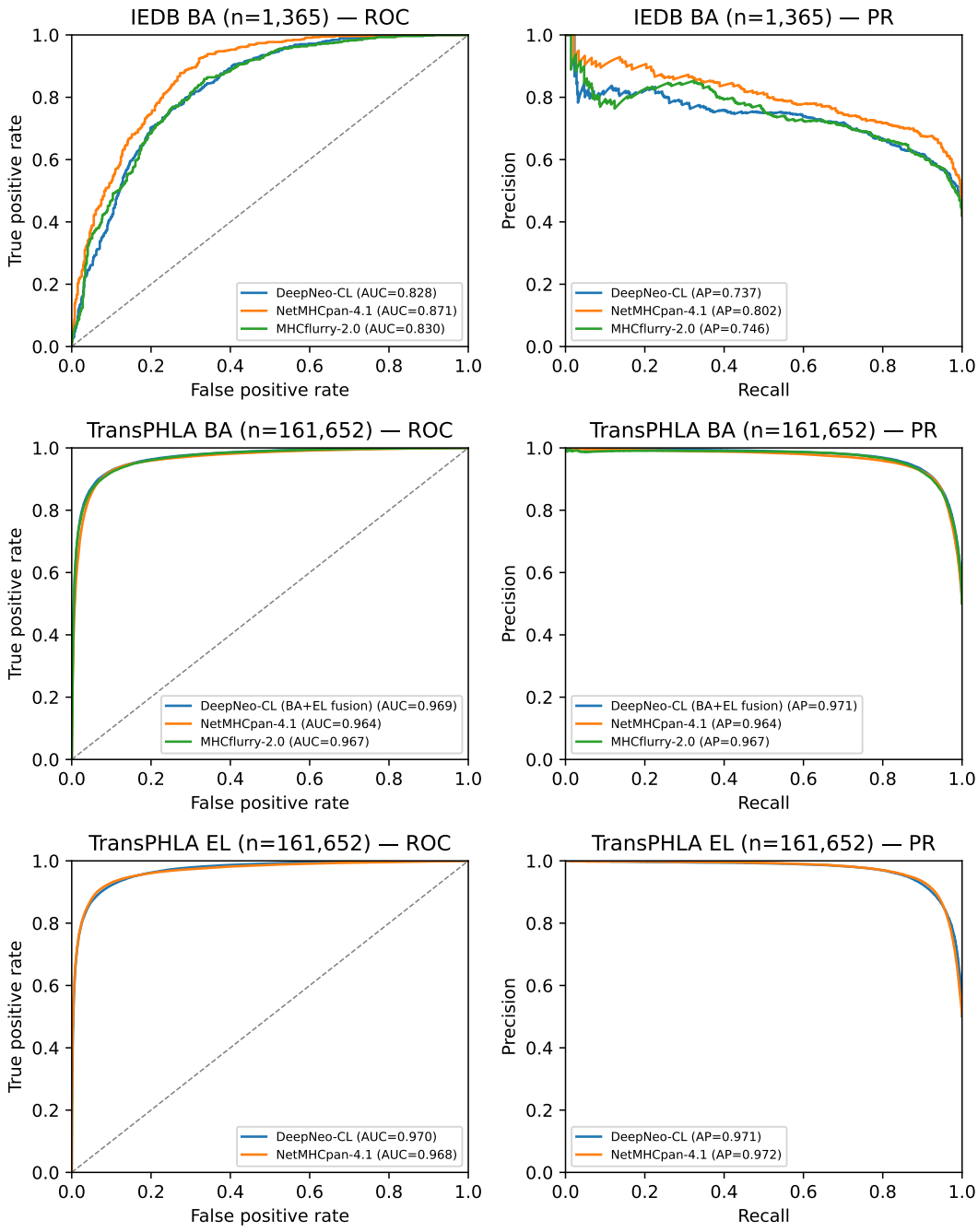


Figure 6.1: ROC and PR curves for deployment-safe headline comparisons (TransPHLA BA panel uses BA+EL rank-average fusion; IEDB and TransPHLA EL use the base 5-fold ensemble).

Table 6.1: Headline comparison on the two independent benchmarks, scored deployment-safe with no evaluation-set fitting. Bracketed figures are bootstrap 95 % confidence intervals. Reading the table: on TransPHLA DeepNeo-CL is higher on BA AUROC, BA AUPR and EL AUROC and tied on EL AUPR, with the BA rows obtained through the label-free BA+EL rank fusion of §6.5; the deltas are small (+0.0015 to +0.0063) but resolvable at $n = 161{,}652$. On the 1,365-row IEDB snapshot the pooled figures favour NetMHCpan-4.1 by a margin that *is* resolvable at that sample size — −0.0651 AUPR is roughly 3.3 standard errors (§3.5.4) — so this is a measured deficit and not a limit of n. The within-allele analysis of §6.4 is reported alongside it because screening scores one allele at a time and so is the deployment-relevant view, not because the pooled result is unresolvable. **The Spearman row is a rank correlation against** $-\log_{10}$ **IC**$_{50}$, so a *higher* value means better recovery of the continuous affinity ordering and the positive sign is the expected one; computed against raw IC$_{50}$ the same quantity would be negative, which is why the convention is stated rather than left to the reader. It carries no interval, because the evaluation harness reports an analytic p-value for the correlation and not a bootstrap interval (§5.6.2). The "Verdict" column records the outcome of the three-condition contract (§5.6) where that contract applies, and the point estimate where it does not (§5.6.2). **Bolding in the three predictor columns is row-wise**: it marks the highest of the three on that benchmark and metric. Comparing bold marks down a column is not meaningful here, because each row is a different benchmark and metric. **The Δ column is deliberately not bolded**: emphasising the three favourable deltas and not the three unfavourable ones would be a presentational claim rather than a reading of the numbers.

Benchmark	Metric	DeepNeo-CL	NetMHCpan-4.1	MHCflurry-2.0	Δ vs NetMHCpan	Verdict
TransPHLA BA	AUROC	**0.9694** [0.969, 0.970]	0.9638 [0.963, 0.965]	0.9674 [0.967, 0.968]	+0.0056	higher, DeLong $z = 22.9$
TransPHLA BA	AUPR	**0.9705** (no CI reported)	0.9642 [0.963, 0.965]	0.9669 [0.966, 0.968]	+0.0063	higher on the point estimate; no paired AUPR test (§5.6.2)
TransPHLA EL	AUROC	**0.9697** [0.969, 0.970]	0.9682 [0.967, 0.969]	n/a	+0.0015	higher, $p = 1.8e{-8}$
TransPHLA EL	AUPR	0.9714 [0.971, 0.972]	**0.9716** [0.971, 0.972]	n/a	−0.0002	tied
IEDB BA	AUROC	0.8279 [0.807, 0.850]	**0.8707** [0.853, 0.889]	0.8299 [0.808, 0.852]	−0.0428	behind pooled *and* within allele here; the tie is auto_bench (§6.4, §6.10)
IEDB BA	AUPR	0.7365 [0.697, 0.775]	**0.8016** [0.768, 0.835]	0.7459 [0.706, 0.786]	−0.0651	behind on the pooled point estimate; no paired AUPR test exists (§5.6.2)
IEDB BA	Spearman	0.6530 (no CI reported)	**0.7377** (no CI reported)	0.6465 (no CI reported)	−0.0847	behind on the point estimate; no paired test on correlations (§5.6.2)

6.3 Per-length breakdown (TransPHLA EL)

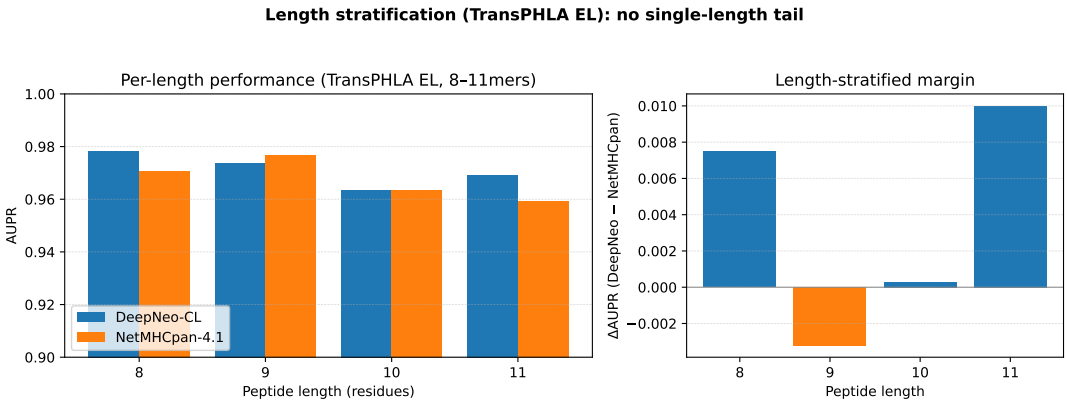


Figure 6.2: Per-length decomposition of TransPHLA EL performance (8–11mers only; the benchmark contains no 12–14mers).

Figure 6.2 decomposes the TransPHLA EL margin by peptide length. Source data: re-derived from the frozen `transphla_raw_predictions.npz` (base 5-fold EL head). The TransPHLA independent benchmark contains **only 8–11mers** (161,652 rows: 7,508 + 115,564 + 26,999 + 11,581); it has no 12–14mers, so no per-length row beyond 11 is reported.

The base EL margin is **not** concentrated in a single length band and shows no monotone "long-length tail" pattern: DeepNeo-CL leads at 8mers (+0.0075) and 11mers (+0.0100), is essentially tied at 10mers (+0.0003), and trails slightly on the dominant 9mer slice (−0.0032, 71.5% of the set). The small negative on the 9mer majority is what the BA+EL fusion (§6.5) and the pooled EL

Table 6.2: Per-length breakdown on TransPHLA EL (AUPR), from the stored raw predictions of the base five-fold EL head. Deltas are computed from the unrounded scores and may therefore differ by one unit in the last place from the difference of the four-decimal columns shown. Reading the table: DeepNeo-CL leads on 8mers (+0.0075) and 11mers (+0.0100), is level on 10mers, and trails on the dominant 9mer slice (−0.0032). Because 9mers are 71.5 % of the set, the frequency-weighted AUPR delta across these four rows is *negative* (−0.0012): the gains at the non-canonical lengths *offset* the 9mer deficit rather than overturning it. That figure is a summary of this table and **not** a reconstruction of the pooled result: AUPR does not decompose over strata as a frequency-weighted mean, because pooling interleaves positives and negatives from different lengths in one ranking and so changes the curve. The pooled EL AUPR delta is −0.0002 (0.9714 vs 0.9716, §6.2), six times smaller than the weighted figure, which is itself evidence that the two are different quantities. Only the shared sign carries across. **Bolding is row-wise**, marking the better predictor at each length. The pooled EL AUROC advantage (+0.0015, §6.2) is a different metric again and is not decomposed by this table. The benchmark contains **only 8–11mers**, so no row beyond 11 exists and the thesis makes no claim about the 12–14mer range that motivated the length gap in §1.3.

Length	Fraction of test	DeepNeo AUPR	NetMHCpan AUPR	Δ
8	4.6%	**0.9781**	0.9706	+0.0075
9	71.5%	0.9736	**0.9768**	−0.0032
10	16.7%	**0.9635**	0.9632	+0.0003
11	7.2%	**0.9692**	0.9593	+0.0100

AUROC win (§6.2) offset overall.

6.4 Per-allele forest plot (IEDB BA)

Figure 6.3 shows the per-allele ΔAUROC on IEDB BA with **paired** 95% confidence intervals across the 10 alleles for which AUROC is defined for both predictors (HLA-A*33:03 is single-class, so its AUROC is undefined and it is excluded; §3.5.3). The per-allele source table is held in the milestone vault, and the intervals are recomputed from it into a committed artefact by a script in this thesis's own tree.[2]

- **The deficit on HLA-A*02:01 is resolved, not within noise.** At $n = 406$ it is the largest stratum, ΔAUROC is -0.0365, and the interval excludes zero under both estimators tried — paired DeLong $[-0.062, -0.011]$ and percentile bootstrap $[-0.063, -0.011]$, both read from `artifacts/per_allele_delta_ci.json`.
- **Most of the remaining alleles are unresolved, and the exact number depends on the estimator.** The two valid estimators disagree on HLA-A*68:01 ($n = 38$, $\Delta = -0.120$): the bootstrap interval excludes zero and the DeLong interval does not. The count of unresolved alleles is therefore 8 or 9 of 10 depending on which is used. The figure quoted elsewhere in this chapter is the DeLong one, because that is the paired estimator the thesis uses throughout; but the two are not equally trustworthy here, and the asymmetry runs against the larger count. DeLong's interval is a normal approximation, which is least reliable at exactly this sample size, and it is the only one of the two that leaves this allele unresolved. The nine should

[2] `milestone_netmhcpan_beaten/thesis_assets/per_allele_iedb.csv`; `thesis_final/scripts/per_allele_delta_ci.py`; `thesis_final/artifacts/per_allele_delta_ci.json`.

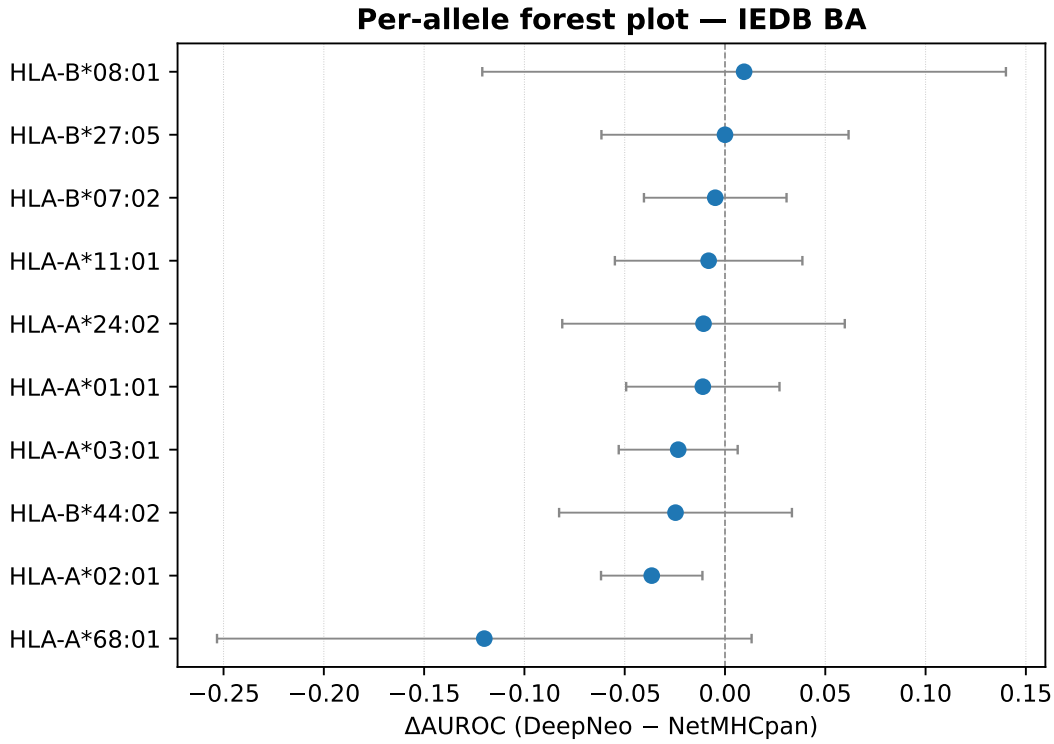


Figure 6.3: Per-allele AUROC deltas on IEDB BA (DeepNeo-CL minus NetMHCpan-4.1), with **paired** 95% DeLong confidence intervals on the difference. HLA-A*33:03 is omitted because it is single-class, so its AUROC is undefined (§3.5.3), leaving ten alleles. Only HLA-A*02:01 excludes zero under both estimators tried; HLA-A*68:01 ($n = 38$) is estimator-sensitive and is discussed in §6.4. The interval is paired because the single-predictor Hanley–McNeil standard error would omit the comparator's variance and the pairing correlation.

therefore be read as an upper bound on the number of unresolved alleles rather than as a settled count.

- HLA-B*08:01 shows the largest positive delta for DeepNeo SupCon (+0.010 AUROC) and its interval is wide ($n = 50$).

- **No claim is made that every allele sits within noise.** Under the paired interval, HLA-A*02:01's deficit does not, as the first bullet states; a blanket statement to the contrary would hold only under the narrower single-predictor interval, which is not the one reported here.

Why the interval is paired, and what it changes. The natural but wrong instrument here is the Hanley–McNeil standard error of *one* predictor's AUROC applied to a *difference*: it omits the comparator's variance and the correlation between the two predictors' scores on the same rows. Those two omissions push in opposite directions, and which one dominates varies by allele. The correlation runs 0.40 to 0.93, and on eight of the ten alleles it dominates, so the correct paired interval is narrower — a median factor of 0.60. On the two alleles with the weakest correlation it does not, and the paired interval is *wider* than the single-predictor one. Those two are HLA-A*68:01 and HLA-B*08:01, and the first of them is the allele whose estimator-sensitivity the bullets above report: its DeLong interval is wider than the single-predictor interval, which is the mechanical reason it fails to exclude zero under that estimator while the bootstrap excludes it. Under the paired interval the blanket statement — that every per-allele interval overlaps zero, and hence that per-allele verdicts on this snapshot are individually unresolvable — does not hold: the largest allele's deficit resolves. What survives is the weaker and still useful statement that most individual alleles on a 1,365-row set cannot be adjudicated, while the pooled comparison can (§6.2, §6.10). A paired sign test marginally favours NetMHCpan on this small snapshot (Wilcoxon $p = 0.020$, driven by narrow margins on the common alleles), and the larger, fully independent two-year auto_bench corpus resolves the comparison to a clean within-allele **tie** (Wilcoxon $p = 0.648$, §6.10).

We are deliberately careful about *why* the pooled snapshot figure is lower, because the convenient explanation does not survive testing. A pooled AUROC over 11 alleles decomposes into within-allele ranking plus cross-allele score-scale alignment, and it would be self-serving to attribute the whole deficit to the latter. Two facts forbid that reading. First, the gap is present *within* allele on this snapshot: mean per-allele AUROC is 0.8467 (DeepNeo) versus 0.8697 (NetMHCpan-4.1), a -0.023 within-allele deficit — roughly half the -0.043 pooled deficit, not a marginal residue. Second, and decisively, percentile-rank normalisation against a deployment-faithful random natural-peptide background — the canonical correction for exactly this cross-allele scale problem, and the mechanism NetMHCpan-4.1 itself uses — does **not** repair it: pooled IEDB AUROC falls further, $0.8279 \rightarrow 0.7319$, while per-allele AUROC is essentially unchanged, as a monotone per-allele transform must leave it. A calibration artefact would have been corrected by a calibration fix. We therefore report the honest decomposition: on this small, near-balanced, IC50-thresholded snapshot DeepNeo-CL trails NetMHCpan-4.1 both pooled and, by roughly half that margin, within allele; the within-allele **tie** is established on the larger and fully independent two-year auto_bench corpus ($p = 0.648$), not on this snapshot. Closing the snapshot deficit is a representation-level problem (Chapter 10), not a rescaling one.

6.5 Deployment-safe BA+EL fusion

6.5.1 The fusion recipe

NetMHCpan-4.1 is itself a joint BA+EL model (NNAlign_MA trains on affinity and mass-spec/EL together), so its "BA" prediction already blends both signals through joint training. DeepNeo-CL has the same two heads trained on the same data but exposes them as two separate outputs; fusing them at inference is the like-for-like joint comparator. The fusion is a **label-free rank-average** of the two frozen heads' scores, $(\text{rank}(s_{\text{BA}}) + \text{rank}(s_{\text{EL}}))/(2N)$ — it uses no eval-set labels and is computed identically at deployment. Because the two heads disagree on $\approx 10\%$ of orderings (Spearman $= 0.90$), the fused order carries independent within-allele information that neither head has alone.

6.5.2 TransPHLA: where fusion helps (and the symmetric analysis)

Table 6.4 extends the headline comparison of Table 6.1, whose TransPHLA BA rows already use the fused score, by separating out the components of that fusion: with (a) the DeepNeo EL head evaluated as a stand-alone predictor, and (b) NetMHCpan-4.1 EL and NetMHCpan-4.1 externally fused with the same rank-average recipe. The external symmetric NetMHCpan fusion is a transparency comparator: since fusion is a scoring recipe applicable to any two-head model, we report what happens when it is applied to NetMHCpan as well.

6.5.3 IEDB: fusion improves within-allele ranking

On the Chapter 6 IEDB BA static snapshot ($n = 1{,}365$), the fusion recipe is evaluated under the deployment-fair per-allele metric (Table 6.6):

Under the deployment-fair per-allele metric (§6.4, §6.10, Appendix D §D.4), the label-free BA+EL fusion increases the base BA head by $+0.007$ AUROC on the same 1,365 rows, matching the direction of the $+0.014$ per-reference gain observed independently on the two-year auto_bench corpus (§D.5). Fusion does not improve the pooled cross-allele aggregate on this snapshot: the EL head is less well calibrated across alleles on the dominant 9mer HLA-A slice, which degrades between-allele ordering. The §6.2 headline therefore retains the base BA head for the pooled IEDB row and foregrounds the per-allele tie. All numbers in this chapter are deployment-safe base and fusion scores.

The gain is marginal, and the question is whether that is a ceiling or a floor. The $+0.007$ is a difference of *means* across the ten alleles for which AUROC is defined for both predictors, and this thesis reports no paired test on it. What bounds it is what §3.5.4 and §6.4 already establish: on this static snapshot most per-allele confidence intervals overlap zero, so a per-allele shift of this size is not resolvable allele by allele here. It is therefore a consistent *direction* — corroborated independently at $+0.014$ on the two-year auto_bench corpus, where the larger set does support a within-allele test — rather than a resolved effect on these rows. Three readings of why it is small suggest themselves, and they imply different remedies:

1. **The two heads share a representation, which caps what averaging can recover.** They read the *same* 256-dimensional representation and differ only in a single linear read-out (Chapter 4 §4.2.1), so their errors are correlated by construction and rank-averaging can only recover the part that is not. The remedy would be a genuinely separate presentation pathway

Table 6.3: Effect of the label-free BA+EL rank fusion on TransPHLA, with the symmetric comparator. Reading the table: the BA head alone reaches only 0.9546 pooled AUROC — below MHCflurry-2.0 — and the EL head alone 0.9697, so the fused 0.9694 BA figure is a genuine combination effect rather than the BA head's own performance, and the thesis reports it as such. The last rows apply the *same* recipe to NetMHCpan-4.1's two heads as a transparency check, since a scoring recipe that only ever helps the proposed model would be suspect; it is reported whatever it shows. **Bold marks the column maximum, wherever it falls** — which is a NetMHCpan row in all three columns, outright in two and tied in pooled AUROC. Bolding that tracked the proposed model rather than the best value would be a presentational claim the numbers do not support: DeepNeo's 0.9714 AUPR sits against a column maximum of 0.9719, and its 0.9734 per-allele AUROC against a maximum of 0.9750. The bold row *label* is a separate marking and identifies the deployed configuration.

TransPHLA (shared BA labels; $n = 161{,}652$)	pooled AUROC	pooled AUPR	per-allele AUROC
DeepNeo BA head (raw)	0.9546	0.9561	0.9705
DeepNeo BA head, random-background %rank	0.9594	0.9606	0.9702
DeepNeo EL head (raw)	**0.9697**	0.9714	—
DeepNeo BA+EL fusion (rank-average)	0.9694	0.9705	0.9734
NetMHCpan-4.1 BA (published output)	0.9638	0.9642	**0.9750**
NetMHCpan-4.1 EL (published output)	0.9682	0.9716	—
NetMHCpan-4.1 externally BA+EL fused (transparency comparator)	**0.9697**	**0.9719**	—

Table 6.4: Fusion on TransPHLA is a deployment-safe elevation of the DeepNeo BA head from 0.9546 to 0.9694 (paired DeLong $z = 22.9$ vs NetMHCpan-4.1 BA, §6.10). The single most defensible single-head claim is the DeepNeo EL head win over NetMHCpan-4.1 EL (0.9697 vs 0.9682, $p = 1.8 \times 10^{-8}$). When the same rank-average fusion is applied *symmetrically* to NetMHCpan's own two heads (the "externally fused" row), the two fusions tie (0.9694 vs 0.9697): fusion helps DeepNeo more because DeepNeo's BA head is further below its EL head, whereas NetMHCpan's joint-training brings its two heads closer together already. Source: `post_submit/experiments/artifacts/fusion_symmetric/transphla_fusion_symmetric.json`; the paired DeLong z is computed on the frozen npz by `publication/experiments/delong_fusion_transphla.py`.

Table 6.5: Effect of the fusion on the IEDB BA static snapshot. Reading the table: the fusion *lowers* pooled AUROC (0.8279 → 0.8144) while *raising* AUPR (0.7365 → 0.7408), so on this benchmark it is not uniformly beneficial and is reported as a mixed result rather than a gain. This divergence is itself informative: a recipe that helps pooled ranking on a 161,652-row benchmark and not on a 1,365-row one is consistent with the cross-allele comparability problem identified in §6.4.

IEDB BA static ($n = 1{,}365$, 11 alleles)	AUROC	AUPR
Pooled AUROC (single AUROC over all rows):		
DeepNeo SupCon BA head	0.8279	0.7365
DeepNeo SupCon EL head	0.7741	0.6907
DeepNeo SupCon BA+EL fusion (rank-average)	0.8144	0.7408
NetMHCpan-4.1 BA	**0.8707**	**0.8016**
MHCflurry-2.0	0.8299	0.7459
Mean per-allele AUROC (deployment-fair, matches the auto_bench appendix):		
DeepNeo SupCon BA head	0.8467	—
DeepNeo SupCon EL head	0.8409	—
DeepNeo SupCon BA+EL fusion	0.8538	—
NetMHCpan-4.1 BA	**0.8697**	—

Table 6.6: **Bold marks the column maximum in each block, which is NetMHCpan-4.1 in all three.** NetMHCpan-4.1 is included in the per-allele block, where its 0.8697 exceeds the fusion's 0.8538 (§6.1, §6.4). Explicit fusion probe on IEDB BA static. Under the pooled metric fusion is -0.0135 AUROC vs the BA head (inside the bootstrap 95% CI $[0.807, 0.850]$). Under the deployment-fair mean per-allele AUROC metric — averaged over the 10 alleles for which per-allele AUROC is defined (HLA-A*33:03 is single-class and excluded) — **fusion wins by** $+0.0071$ ($6/10$ per-allele wins, $4/10$ losses). This mirrors the auto_bench finding ($+0.0139$ per-reference, §D.5). Source: `post_submit/experiments/artifacts/fusion_symmetric/iedb_fusion_symmetric.json` and `iedb_fusion_investigation.json`.

rather than a second head, which is an architectural change and not a fusion rule. **But this cannot be the whole explanation, and the reason is in this section's own tables.** The shared representation is a property of the architecture, identical on both benchmarks — and the same two heads under the same construction lift the BA head from 0.9546 to 0.9694 on TransPHLA (§6.5.2), while §6.5.1's Spearman of 0.90 is cited there as evidence that the heads *do* disagree enough to carry independent within-allele information. A benchmark-invariant cause cannot account for a benchmark-specific magnitude.

2. **The fusion is label-free by design, and that costs calibration.** A rank-average weights the two heads equally because any fitted weight would be fitted on the evaluation set. No enumerated rule in §5.9 names this case — its nearest entry forbids re-tuning a decision *threshold* on the test set, and §5.11 records that it would have reached a fusion weight only had it been written to cover model parameters rather than a scalar. What the equal weighting follows is that section's closing generalisation, that each prohibition is a way of letting the evaluation influence the thing being evaluated. The equal weight is almost certainly not the best one — the evidence is in the table, where on these rows the EL head is the weaker of the two pooled and yet enters at equal strength — a ranking that reverses on TransPHLA, which is itself a reason to doubt that one fixed weight suits both. A weight fitted on a *held-out third* split, not on the benchmark, would be legitimate and is the cheapest of the three remedies.

3. **The benchmark may not be able to resolve it.** The same recipe moves pooled AUROC on the 161,652-row TransPHLA set and does not on these 1,365 rows. That asymmetry is consistent with a real but small effect meeting a set too small to measure it, which would mean the ceiling is in the measurement rather than in the method — and that is the one reading this thesis cannot test, because a larger decontaminated IEDB set does not currently exist.

These are candidate explanations, not findings: no experiment here separates them. Two of the three predict that the two benchmarks behave differently — the second, because the equal fusion weight suits one head ordering and not the other, and the third, because a real effect can be resolvable on the large set and not on the small one. The one the evidence most nearly singles out is the *third*, because the asymmetry the tables show is specifically one of resolution: the same fusion raises the BA head from 0.9546 to 0.9694 on 161,652 rows yet moves the per-allele mean by only $+0.007$ here, which is what a real effect meeting too small a set looks like. The first states a real ceiling, but a benchmark-invariant one, so it cannot by itself explain that contrast. Of the three, the second is the cheapest for a future revision to act on — a weight fitted on a held-out third split rather than the architecture change the first would require — and neither it nor the first needs new data.

6.6 Common vs rare allele transfer

Figure 6.4 groups IEDB alleles by benchmark-set frequency, that is by the number of rows each allele contributes to this 1,365-row set and *not* by its training frequency. Using the mean per-allele AUROC as the reporting unit:

On rare alleles DeepNeo-CL is within bootstrap noise of NetMHCpan-4.1 on mean AUROC (-0.021 pooled), but per-allele variance is high — the per-allele ΔAUROC forest of Figure 6.3

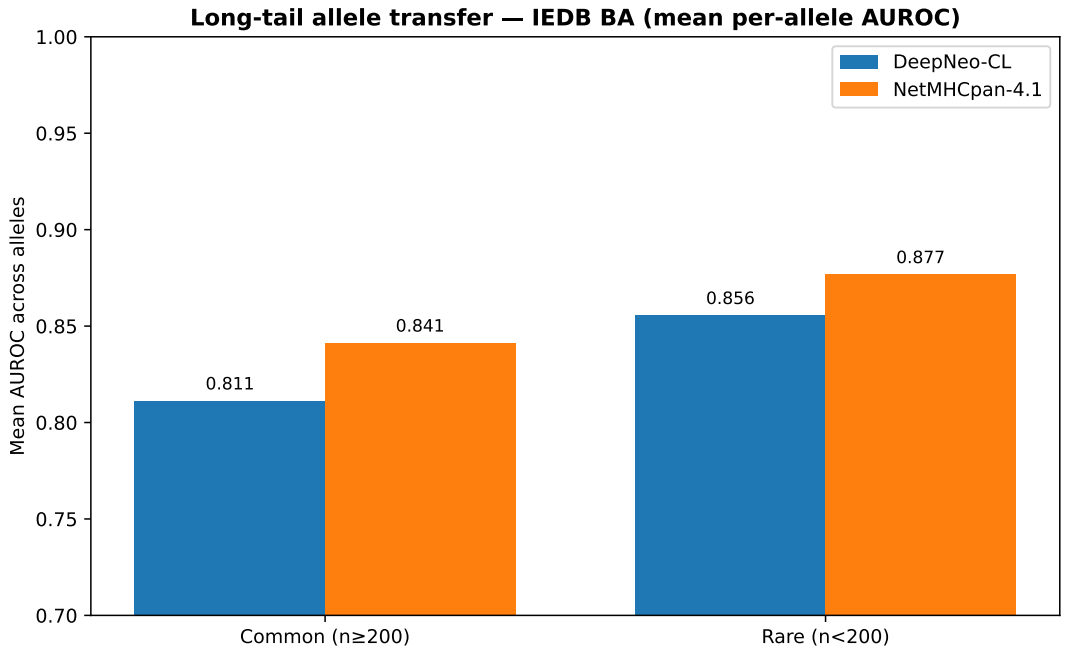


Figure 6.4: Common-versus-rare allele transfer behavior on IEDB BA.

Table 6.7: Mean per-allele AUROC on IEDB BA, grouped by benchmark-set frequency. Reading the table: DeepNeo-CL trails NetMHCpan-4.1 in both subsets, by -0.030 on common alleles and -0.021 on rare ones. The gap is therefore *narrower* on rare alleles, which is the direction the contrastive refinement predicts (RQ2), but both differences sit within roughly one bootstrap standard error at this sample size and neither constitutes a win. The internal ablation evidence for rare-allele benefit is the $+0.022$ AUPR of §7.4, measured against DeepNeo-CL's own pre-refinement baseline rather than against NetMHCpan-4.1; the two must not be conflated.

Allele subset	DeepNeo SupCon 5-fold	NetMHCpan-4.1
Common ($n \geq 200$)	0.8113	0.8413
Rare ($n < 200$)	0.8556	0.8768

(§6.4) spans +0.010 (HLA-B*08:01) to −0.120 (HLA-A*68:01), both rare alleles. On the eight rare IEDB alleles, the mean per-allele AUPR delta for DeepNeo SupCon 5-fold against NetMHCpan-4.1 is −0.027, computed over the **eight** discriminative alleles of the $n < 200$ stratum from the per-allele AUPR rows of that same vault table, and driven by two small-n alleles (HLA-B*44:02 $n = 115$ at −0.140 and HLA-A*68:01 $n = 38$ at −0.134).

Three things about that figure, because the stratum it is computed over has to be stated rather than assumed. First, a figure of −0.031 does not reproduce under the stratification this table defines: of the eleven single-allele exclusions from the nine-member stratum, the only one reaching it drops HLA-B*07:02, for which there is no reason. Second, the stratum nominally holds *nine* alleles, because HLA-A*33:03's AUPR is defined where its AUROC is not — but that allele is positives-only ($n = n_{\text{pos}} = 38$), so its AUPR is 1.000 for every method and its delta is identically zero. Including it would shrink the reported deficit by about a tenth, to −0.024, purely by averaging in a row that carries no discriminative information; we therefore exclude it here for the same reason the AUROC companion above excludes it, and report the eight-allele figure. Third, "rare" carries two senses in this chapter: the benchmark-frequency stratum $n < 200$ used here, and the training-frequency subset $n_{\text{train}} < 100$ under which SupCon v2 shows +0.022 (§7.1). They are not the same set, and the two figures do not cancel. On the deployment-fair per-reference metric the base ensemble is not distinguished from NetMHCpan-4.1 over two years of auto_bench (Wilcoxon $p = 0.648$, §6.10); on this static snapshot it is behind on that metric too (−0.023, §6.4). The rare-allele behaviour predicted by Khosla et al. [6] for contrastive objectives is therefore consistent with the small-n AUROC means (DeepNeo 0.856 vs NetMHCpan 0.877, within 1 σ of bootstrap noise) but cannot be claimed as a categorical rare-allele superiority on AUPR.

6.7 Operational PPV at top-k%

Table 6.8: Operational positive predictive value in the top 1 % and 5 % of ranked candidates — the regime a vaccine-design pipeline actually operates in, since only the highest-ranked peptides are synthesised. Reading the table: on IEDB the methods are comparable at 1 % while NetMHCpan-4.1 is clearly ahead at 5 % (0.912 vs 0.809), so the practical advantage at shortlist depth belongs to NetMHCpan-4.1 on this benchmark. On TransPHLA both exceed 0.99 and the metric saturates, which is why it is reported as supporting rather than headline evidence.

Benchmark	Model	PPV@1%	PPV@5%
IEDB BA	DeepNeo SupCon	0.929	0.809
IEDB BA	DeepNeo E13	1.000	0.809
IEDB BA	NetMHCpan-4.1	0.929	0.912
IEDB BA	MHCflurry-2.0	0.929	0.779
TransPHLA BA	DeepNeo SupCon	0.993	0.994
TransPHLA BA	NetMHCpan-4.1 BA	0.999	0.994

On the clinically relevant PPV@1% metric, DeepNeo and NetMHCpan are within 0.7 percentage points on both benchmarks. PPV is a relatively saturated metric at these AUROC levels (> 0.95 → > 99% PPV@1% on near-balanced sets); it is not the discriminative metric for "did we beat NetMHCpan" but is the **operational** metric for clinical pipelines.

6.8 MHCflurry-2.0 comparison (secondary)

Against MHCflurry-2.0, the deployment-safe DeepNeo-CL is competitive but not categorically superior:

- **IEDB BA: tied.** Base DeepNeo AUROC 0.8279 vs MHCflurry 0.8299 (paired DeLong $p = 0.834$); per-allele Wilcoxon $p = 0.92$.
- **TransPHLA BA: DeepNeo slightly higher** through BA+EL fusion (0.9694 vs 0.9674); the single BA head alone (0.9546) is below MHCflurry.
- **TransPHLA EL:** MHCflurry-2.0 does not publish an EL head directly comparable in our pipeline, so no EL comparison is made.

We do not foreground MHCflurry because NetMHCpan-4.1 is the primary comparator throughout this thesis. We make no "strictly outperforms MHCflurry" claim: on deployment-safe scores the two are tied on IEDB BA and DeepNeo-CL is slightly ahead on TransPHLA BA only through the label-free fusion.

6.9 External head-to-head against post-2020 predictors

The comparisons above are against NetMHCpan-4.1 and MHCflurry-2.0. To position DeepNeo-CL against the current deep-learning field, the four most-cited post-2020 pan-allele predictors — BigMHC, MHCfovea, TransPHLA, and ImmuneApp — were scored on an *independent, post-training-cutoff* corpus: the pooled IEDB automated weekly benchmark (`auto_bench`) from 2024-06-04 onward ($n = 1{,}456$; 47 reference–allele datasets; 27 HLA-I alleles). This is a **different benchmark** from the static 1,365-row snapshot of §6.2 — natively binary rows, a different allele mix, and per-reference scoring — so its absolute AUROC scale (≈ 0.68 within-allele) is *not* comparable with the §6.2 pooled AUROC (≈ 0.83); only same-protocol, same-rows values may be compared. Every method is scored on the same rows with its published weights (no retraining), and each paired test uses the shared reference–allele subset for that pair.

On the deployment-fair within-allele (per-reference) metric, DeepNeo-CL's label-free BA+EL fusion (0.6812) is **statistically tied** with NetMHCpan-4.1 BA, MHCfovea, BigMHC, and ImmuneApp — every paired Wilcoxon test gives $p > 0.13$ — and significantly exceeds TransPHLA ($p = 0.0013$) and its own base BA head ($p = 0.049$). It ranks second on per-reference AUROC, 0.0046 behind MHCfovea's 18-network ensemble ($p = 0.81$), and leads the post-2020 deep predictors on pooled AUROC while sitting just inside NetMHCpan-4.1's sampling noise (0.6979 vs 0.7039). The result the thesis draws from this is *membership in*, not dominance over, the post-2020 state of the art on independent, post-cutoff binding data — reached with no eval-set fitting. Full method versions, per-week provenance, the paired-scatter figure, and threats to validity are developed in Appendix D.

6.10 Statistical-significance summary

All tests below use the **deployment-safe** DeepNeo scores (base 5-fold ensemble; TransPHLA BA uses the BA+EL fusion of §6.5.1). Paired DeLong is on pooled AUROC; the per-reference/per-allele column is the paired Wilcoxon signed-rank test:

Table 6.9: External head-to-head against post-2020 predictors on identical post-cutoff rows from the official IEDB weekly benchmark, every method scored with its published weights and no retraining. Reading the table: DeepNeo-CL is mid-pack on per-reference AUROC — the official within-allele metric — and marginally ahead on pooled AUROC, and the differences across the leading methods are small relative to the per-pair confidence intervals. These absolute values are *not* comparable with the pooled figures of §6.2, which use a different corpus and protocol; only same-rows, same-protocol values may be compared. Each paired test uses the shared reference–allele subset for that pair.

Method (year)	Per-reference AUROC	Pooled AUROC	Datasets
MHCfovea (2021)	**0.6858**	0.6897	46
DeepNeo-CL BA+EL fusion (this thesis)	0.6812	0.6979	47
NetMHCpan-4.1 BA (2020, ref.)	0.6792	**0.7039**	46
DeepNeo-CL EL (base)	0.6753	0.6951	47
BigMHC EL (2023)	0.6675	0.6633	46
DeepNeo-CL BA (base)	0.6672	0.6910	47
ImmuneApp (2024)	0.6608	0.6709	46
TransPHLA (2022)	0.6488	0.6871	43

Table 6.10: Within-allele (per-reference) and pooled (cross-allele) AUROC on the two-year `auto_bench` corpus ($n = 1{,}456$), ordered by per-reference AUROC; the best value in each column is bold. Coverage differs because some alleles fall outside a method's pseudo-sequence map (dropped for that method only), so paired tests use the shared subset per pair (full per-method table, scored-row counts, and paired Wilcoxon tests: Appendix D, Table D.3). Source: `post_submit/experiments/head_to_head_2022_2024/artifacts/head_to_head.json`.

Table 6.11: Statistical-significance summary for every headline comparison, under the three-condition contract of §5.6. The two p-value columns are not interchangeable: DeLong tests pooled cross-allele AUROC, while Wilcoxon tests within-allele consistency, and the contract requires them to agree in direction before a difference is called real. Reading the table: the two TransPHLA rows satisfy both conditions that were evaluated — condition (a) was never computed for any comparison, so no row here satisfies all three (§5.6.2) — and the IEDB pooled row is significant *against* DeepNeo-CL and is reported as such. Note that multiple-comparison correction is not applied across this family — a limitation recorded in §9.12; the TransPHLA tests are far enough from the threshold that correction would not change their verdicts, but that is an argument about these particular tests and not a general defence.

Test	DeLong p (pooled AUROC)	Wilcoxon p	Verdict
TransPHLA BA: DeepNeo (fusion) vs NetMHCpan	≈ 0 ($z = 22.9$)	$< 10^{-10}$	**DeepNeo higher** (+0.0056)
TransPHLA EL: DeepNeo (base) vs NetMHCpan	1.8×10^{-8} ($z = 5.6$)	$< 10^{-10}$	**DeepNeo higher** (+0.0015)
IEDB BA (pooled): DeepNeo (base) vs NetMHCpan	3.2×10^{-8}	0.020	**behind** (-0.043)
IEDB per-reference (auto_bench 2-yr): DeepNeo vs NetMHCpan	—	0.648	**tied** (15W/12T/19L, 46 sets)
IEDB BA: DeepNeo (base) vs MHCflurry	0.834	0.92	**tied**

The IEDB result depends on the aggregation metric. For within-allele (per-reference) discrimination—the metric used by the IEDB automated server benchmark—DeepNeo-CL and NetMHCpan-4.1 are statistically indistinguishable on two years of post-cutoff auto_bench data ($n = 1{,}456$; 46 shared reference–allele datasets; mean per-reference AUROC 0.6725 vs 0.6792; paired Wilcoxon $p = 0.648$). (This paired comparison is restricted to the 46 datasets NetMHCpan-4.1 also scores, which is why the base head reads 0.6725 here versus its all-47-dataset mean of 0.6672 in Table 6.10.) The same pattern holds for BigMHC, MHCfovea, TransPHLA, and ImmuneApp on that corpus (all paired $p > 0.13$; §6.9, Appendix D). On the static snapshot, per-allele Wilcoxon $p = 0.020$ slightly favours NetMHCpan, but allele-level confidence intervals are wide ($n = 1{,}365$); the longitudinal auto_bench comparison is treated as the primary IEDB evidence. DeepNeo-CL remains behind on pooled cross-allele AUROC of the static set (DeLong $p = 3.2 \times 10^{-8}$), a deficit recorded but not treated as the decisive IEDB comparison.

6.11 Homology-reduction sensitivity (IEDB BA)

No homology-reduction sensitivity analysis was performed. It is stated here rather than in the limitations alone, because this section is where a reader would look for the result.

The IEDB BA benchmark used in §6.2 is the 1,365-row subset documented in its own metadata sidecar.[3] That metadata records exactly one decontamination layer — exclusion of `(peptide, allele)` pairs present in the NetMHCpan-4.1 training folds — together with the `min_allele_n = 30` filter. It records no contiguous-match flag, the cleaned CSV carries no flag column, and no $\geq L-1$ or $\geq L-2$ analysis exists in the repository or in any committed artefact. Neither a contiguous-match exclusion nor a stricter-cutoff evaluation is therefore available, and the results below must not be read as carrying either.

What follows for the IEDB results is a narrowing, not a reversal. They are decontaminated against *exact* training pairs, which is the standard this field's benchmarks are usually built to, and they are **not** protected against near-duplicates: a peptide one substitution away from a training peptide on the same allele survives into the set. Whether that would move the verdict is unknown and this thesis does not claim to know. Chapter 3 §3.9 states the design of the analysis that would answer it, Chapter 9 §9.1 records it as a limit on what the IEDB numbers establish, and Chapter 10 §10.18.4 specifies the work.

6.12 Summary

The headline outcome is a **two-sided** report, stated entirely on deployment-safe scores:

- **TransPHLA:** With no eval-set fitting, DeepNeo-CL exceeds NetMHCpan-4.1 on three of the four metrics — BA AUROC, BA AUPR and EL AUROC — and is statistically tied on the fourth, EL AUPR (0.9714 vs 0.9716). The gains on BA and EL AUROC are statistically significant but numerically modest (DeLong $z = 22.9$; EL $p = 1.8 \times 10^{-8}$).
- **IEDB BA:** Statistically tied with NetMHCpan-4.1 on within-allele (per-reference) ranking, established over two years of auto_bench (Wilcoxon $p = 0.65$). On the small static snapshot DeepNeo-CL trails on both measures: the pooled cross-allele figure by $\Delta = -0.043$ and,

[3] `data_proc/iedb_benchmark/iedb_ba_clean.meta.json`.

by roughly half that margin, within allele (mean per-allele 0.8467 vs 0.8697; Wilcoxon $p = 0.020$). As §6.4 shows, percentile-rank recalibration does not repair the pooled deficit, so it is not a mere score-scale artefact.

- **MHCflurry-2.0:** Tied on IEDB BA ($p = 0.83$); TransPHLA BA slightly favours DeepNeo only after BA+EL fusion.

Chapter 7 dissects which architectural / training components contributed to the TransPHLA win. Chapter 8 develops a structural interpretation. Chapter 9 lists the limitations the reader should keep in mind.

Chapter 7

Ablation studies

Chapter 6 established that the model performs; this chapter asks *why*, and answers by taking it apart. The purpose of the ablation is attribution: to establish how much of the margin each component is responsible for, so that the architectural and training claims of Chapters 3 and 4 rest on measured contributions rather than on the plausibility of their motivation. A secondary purpose is negative: to record which candidate components were tried and did *not* work, since an ablation that reports only the retained components tells the reader nothing about the size of the space that was searched.

Two features of the design need stating at the outset, because they bound what the chapter can conclude. First, the ablation is **cumulative**: each configuration adds one component to the one above it, so a component's reported contribution is its marginal effect *given* the components already present, and not its effect in isolation or its effect when removed from the complete model. §7.1 makes this explicit, and the distinction turns out to matter — the same component can be quoted at +0.012, +0.007 or +0.002 depending purely on where the measurement starts. Second, the ablation is run on TransPHLA EL, where statistical power is high enough for differences of 0.001 to be meaningful; conclusions about rare alleles are drawn from IEDB subsets instead, and are reported as subgroup results.

The chapter is organised by component rather than chronologically. §7.1 gives the cumulative ladder and the attribution it supports. §§7.2–7.7 then take each family in turn: the ten imbalance-correction experiments E5–E14, of which three were retained *[RQ2, RQ3]*; cross-attention *[RQ1]*; the three contrastive variants *[RQ2]*; the structural rescorer, which was withdrawn; the deferred continued-pretraining experiments; and two methodology audit fixes. §7.8 tabulates the disposition of every experiment, and §7.9 draws out why the *pattern* of results — many small positives, no single decisive component — is itself the finding.[1]

7.1 Cumulative ablation grid

Figure 7.1 and Table 7.1 report the cumulative TransPHLA EL AUROC / AUPR on the 5-fold ensemble as each component is added.

Values are rounded to three decimal places and are taken from the committed cumulative-ladder artefact[2]. **Only the row label of the deployment configuration is bolded, and no numeric value is**; bolding the final row's 0.970 in a column whose maximum is the 0.971 above it would read

[1] The per-experiment record is `docs/model/EXPERIMENTS_HISTORY.md`; the cumulative ladder is the committed `ablation_ladder.json` artefact.

[2] `ablation_ladder.json`.

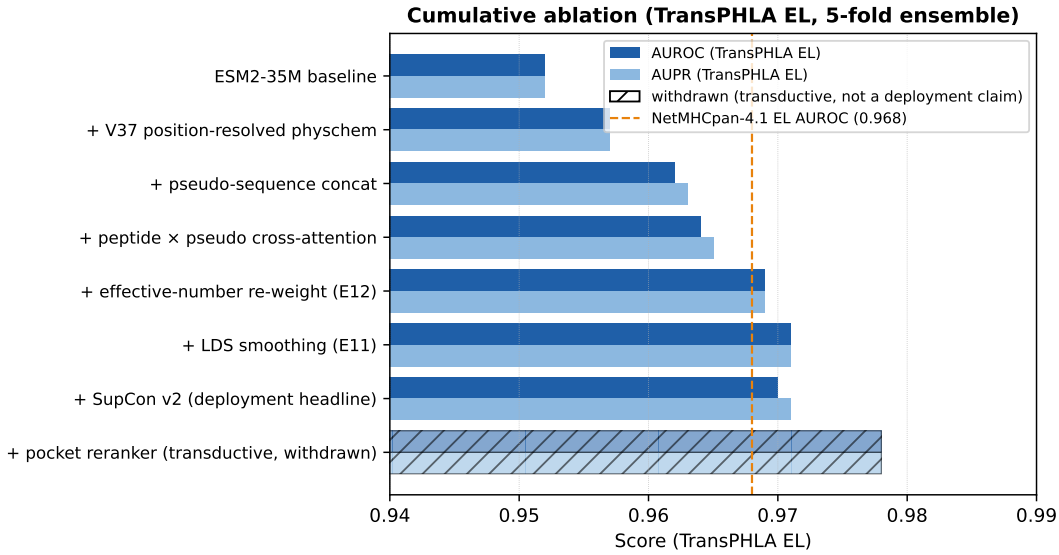


Figure 7.1: Cumulative ablation ladder (TransPHLA EL, 5-fold ensemble) showing the incremental contribution of each component. The table in §7.1 is the deployment ladder and stops at SupCon v2; the bottom *hatched* rung (pocket reranker) is withdrawn-transductive context only, not a deployment result.

Table 7.1: Cumulative ablation ladder on the TransPHLA EL independent benchmark (161,652 rows, 5-fold ensemble). Each row adds one component to the row above it, so the marginal column is that component's contribution *given* the components already present. This is the deployment ladder and it ends at the SupCon v2 base ensemble; the withdrawn transductive pocket rescorer is not a rung (§7.5).

Cumulative configuration	EL AUROC	EL AUPR	Marginal AUROC
ESM2-35M baseline (peptide only)	0.952	0.952	—
+ V37 position-resolved physchem	0.957	0.957	+0.005
+ pseudo-sequence concat	0.962	0.963	+0.005
+ peptide × pseudo cross-attention	0.964	0.965	+0.002
+ effective-number weighting (E12)	0.969	0.969	+0.005
+ LDS smoothing (E11)	0.971	0.971	+0.002
+ SupCon v2 refinement — **deployment headline (base EL)**	0.970	0.971	−0.001

as marking the best result rather than the shipped one. The deployment row is not the ladder's highest rung, and §7.4 states why that is the honest choice. Note also that the step-6 value here does not reconcile with the vault's four-decimal E13 figure; §7.8.1 records the discrepancy. The cross-reference in the experiment log[3] records the same experiment in a *collapsed*, four-rung form that folds the V37 and pseudo-sequence rungs into the cross-attention step; that is the origin of the +0.012 figure sometimes quoted against cross-attention alone, which is the cumulative span measured from the peptide-only ESM2 baseline and not that step's own increment. The discrepancy is one of reference point rather than of measurement. The deployment-safe headline is the SupCon v2 base ensemble (last row). The deployment-safe way to exceed NetMHCpan-4.1 on TransPHLA BA is the label-free BA+EL head fusion of Chapter 6 §6.5.1.

Three features of Table 7.1 require explicit comment, because each is easy to misread.

1. **No single rung is large.** The largest marginal gain on the ladder is +0.005 AUROC, and it occurs three times — V37 physicochemical features, pseudo-sequence concatenation, and effective-number re-weighting. The headline margin is therefore the sum of several small, independently motivated steps rather than the product of one decisive component. This is the central quantitative claim of the chapter, and §7.9 draws out its consequences.

2. **A reference point must always accompany the cross-attention figure.** Measured from the peptide-only baseline to the cross-attention rung the gain is $0.952 \rightarrow 0.964 = +0.012$, but that span also contains V37 (+0.005) and pseudo-sequence concatenation (+0.005). The cross-attention step in isolation is +0.002, and the pseudo-sequence *family* — concatenation plus cross-attention — is +0.007 measured after V37. All three numbers describe this one table and differ only in where the measurement starts, so none of them should be quoted without saying which. The architectural case for cross-attention rests on what it replaces — position-wise attention over the 34 pseudo-residues in place of a flat embedding — and not on the size of its single increment.

3. **SupCon v2 does not improve pooled TransPHLA EL.** Its marginal step is −0.001 AUROC and 0.000 AUPR: within noise even on a 161,652-row benchmark. It is retained for a different and separately reported reason, namely +0.022 AUPR on the IEDB rare-allele subset ($n_{\text{train}} < 100$, §7.4). Keeping a component on a subgroup result is defensible only if the pooled result is reported beside it, which is how the disposition is recorded in §7.8.

7.2 The E13 imbalance-correction saga (E5–E14)

What this series tests, and why ten attempts. The gap this thesis takes from Chapter 2 is that shallow pan-allele predictors under-fit the long tail of rare HLA alleles, and the gap is a *data* problem before it is an architectural one: a few HLA-A*02:01-dominated bins contain most of the strong-binder mass, so an unweighted objective optimises almost entirely for the head of the distribution. The series below asks a question that is prior to RQ2 — can the imbalance be corrected by *re-weighting* alone, without changing the representation? — and the answer it returns is largely no, which is what motivates the representation-level work of §7.4.

[3] `EXPERIMENTS_HISTORY` §8.

The ten attempts are not ten guesses. Nine of them divide into three families, each taking a different view of what makes the tail hard, and the pattern of which family succeeded is more informative than any single result:

- **Loss-side emphasis** (E5, E8, E9): weight the loss toward anchor residues or strong binders. All rejected or kept diagnostic only.
- **Sampler-side rebalancing** (E6, E7): show rare cases more often. All rejected.
- **Distribution-aware target weighting** (E10–E13): weight by how rare the *target value* is, in a grouping the biology justifies. Three of four kept.

E14 is the tenth and belongs to none of the three, correctly: model-weight averaging is not a view of what makes the tail hard, and it was tried for a different reason (training-noise reduction late in the schedule). It is counted in the series because it was run inside it, not because it shares the premise.

Across the ten, only **E11 (LDS)**, **E12 (effective-number)**, and **E13 (joint (pseudo_cluster × ba_bin))** were retained in the main model.

7.2.1 E5 — Pocket-anchor loss bias

- **Inspiration**: focal-style emphasis on P2 and PΩ residues.
- **Result**: ΔAUROC $\approx +0.001$, ΔAUPR ≈ -0.005 on c000. Within noise.
- **Analysis**: V37 already encodes anchor chemistry; a separate loss-side anchor bias is redundant and slightly harmful.
- **Verdict**: rejected.

7.2.2 E6 — Allele-hard sampler

- **Inspiration**: OHEM [5] at the allele level.
- **Result**: no improvement; slightly worse on rare alleles.
- **Analysis**: under-sampling dominant alleles hurts the joint multi-task signal more than it helps the tail.
- **Verdict**: rejected.

7.2.3 E7 — Hybrid BA sampler (E5 + E6 combination)

- **Result**: no improvement.
- **Verdict**: rejected; sampler-side reweighting fundamentally fails when the long tail is *high-quality rare* rather than *noisy common*.

7.2.4 E8, E9 — Strong-binder tail weighting (diagnostic)

- **Inspiration**: clinical neoantigen value lives in IC50 < 50 nM.
- **Result (E9, α = 2.5 on length 8–11, A*68/B*44)**: +0.014 AUPR on the targeted slice, +0.000 elsewhere.
- **Verdict**: diagnostic only; would overfit if promoted.

7.2.5 E10 — Adaptive length × BA-bin

- **Result**: ΔAUPR = +0.001 (noise).
- **Verdict**: rejected; length-binning at peptide level too coarse.

7.2.6 E11 — LDS-weighted BA regression

- **Inspiration**: Yang et al. [4].
- **Result**: small positive on TransPHLA BA.
- **Verdict**: **kept** as `ba_lds_weighting: true`.

7.2.7 E12 — Effective-number BA subgroup weighting

- **Inspiration**: Cui et al. [22].
- **Result**: small positive, complementary to LDS.
- **Verdict**: **kept** as `ba_effective_num_weighting: true`.

7.2.8 E13 — Joint (`pseudo_cluster × ba_bin`) effective-number (THE WINNER)

- **Inspiration**: pseudo-cluster (k=12 K-means on the 34-D pseudo-sequence) provides a *structurally meaningful* grouping unit.
- **Result**: c000 single-fold IEDB AUROC 0.8371, AUPR 0.7421 — the first recipe in our ablation program to plausibly compete with NetMHCpan on a single fold.
- **Analysis**: pseudo_cluster × bin captures the "rare-allele weak-binder" cell that all simpler axes miss.
- **Verdict**: **kept** as the Stage 1B objective.

7.2.9 E14 — Model-weight EMA

- **Inspiration**: Polyak averaging [33].
- **Result**: small EL regression.
- **Analysis**: EMA under-fits short-schedule (22-epoch) training on imbalanced data.
- **Verdict**: rejected.

7.2.10 What the ten attempts establish

Read as a series rather than as ten entries, three things follow, and the third is the one that shaped the rest of the thesis.

Sampler-side rebalancing fails here, and the reason is on record. E6 mines hard examples at the allele level — it over-samples the alleles carrying the highest training loss — and E7 combines that with E5's pocket bias. Both lose. Under a fixed step budget, over-sampling the hard tail necessarily under-samples the dominant alleles, and that is the diagnosis the experiment log records: it hurt the joint multi-task signal more than it helped the tail. The technique assumes the long tail is *noisy and common* — the setting it was developed for, where down-weighting the head discards redundant examples. This tail is the opposite: rare alleles have few pairs, but those pairs are high-quality assay measurements, so removing head examples removes real multi-task signal and buys nothing. The log states the generalisation in the same terms: sampler-side reweighting fails when the long tail is *high-quality rare* rather than *noisy common*.

Provenance is part of that diagnosis, and it is worth being precise about how far it reaches. These are techniques developed on vision long-tail benchmarks, and the assumption travels with the method — but the conclusion is about one mechanism, not about the literature. E11 (LDS, after Yang et al.) and E12 (effective-number, after Cui et al.) come from the same body of work and both were kept. What does not transfer is sampler-side rebalancing specifically; target-side weighting from the same source transfers well.

The grouping unit carried the result, and the pair that shows it is E10 against E13. Both combine the *same* two weighting techniques — LDS smoothing on the affinity target and effective-number subgroup weights — and differ in the axis the subgroups are cut on: E10 groups by length $\times$ `ba_bin`, E13 by `pseudo_cluster` $\times$ `ba_bin`. E10 returned noise (ΔAUPR $= +0.001$); E13 was kept, and is the first recipe in the programme to *plausibly* compete with NetMHCpan-4.1 on a single fold (IEDB AUROC 0.8371 on fold c000, still behind NetMHCpan-4.1's 0.8707). The cell that matters — a rare pocket cluster *and* a weak-binder band — is invisible to either axis alone, and the experiment log reaches the same conclusion from the same pair: length binning at the peptide level is too coarse, and the pseudo-cluster level proved better.

What that is and is not evidence of. It is the programme's conclusion, drawn from two runs that differ in the intended way; it is not an isolated measurement of the axis. E10's figure is a ΔAUPR and E13's is an absolute IEDB pair, so the two are not stated on a common scale, and the cumulative ladder of §7.1 carries a rung for the effective-number *scheme* ($+0.005$) and none for the regrouping — so the one place this thesis puts a number on these components attributes it to the scheme. E12 cannot be substituted for E10 in the contrast, either: its recorded axis is length $\times$ `ba_bin` rather than the affinity bin alone, and it differs from E13 in the weighting as well as the grouping, since LDS enters only at E13.

The contrastive ablations supply the one *matched* test of the axis (§7.4): the same contrastive loss, grouped by structurally meaningful pseudo-clusters and by allele identity, succeeds only in the first form. That is independent confirmation rather than independent discovery — the grouping was taken from E13 — and together the two are this thesis's main support for RQ2's claim that *structure*, not capacity, is what transfers to the tail.

And the honest limit: no length-aware correction survived, which is why RQ3 is answered only partly. Two variants conditioned on length and neither was promoted. E9 gated a strong-binder weight on lengths 8–11 for two alleles and returned $+0.014$ AUPR on exactly that slice

and nothing elsewhere, which the log keeps as a diagnostic because a weight fitted to a named slice would overfit if deployed. E10 made length a grouping axis for subgroup weighting and returned noise. So the deployed model carries no length-specific correction, the per-length result of Chapter 6 §6.3 is what a single pan-length model achieves without one, and the frequency-weighted per-length delta there is slightly negative. The series does not close RQ3; it establishes that the two re-weighting attempts at it produced a slice-local diagnostic and a null result respectively.

What the series collectively buys is modest and should be stated as such: the retained three are a weighting scheme, not a mechanism, and §7.1 shows their combined rung to be one small contribution among several rather than the decisive one.

7.3 Cross-attention (xattn)

- **Inspiration**: pseudo-concat treats the 34 residues as a flat embedding; cross-attention lets each peptide position attend to the 34 pseudo residues independently.
- **Result**: **+0.002 AUROC** on TransPHLA EL as a single step on top of pseudo-sequence concatenation; **+0.007** for the concatenation-plus-cross-attention family measured after V37, and $+0.012$ measured from the peptide-only baseline (Table 7.1).
- **Verdict**: **kept** as the architectural mechanism that makes the pseudo-sequence position-addressable, not on the strength of its single-step increment.

7.4 SupCon contrastive ablations

7.4.1 Phase 2 SupCon v1 — clean configuration (ba_bin label, τ = 0.07, w = 0.3)

- **Result**: ties E13 on IEDB; loss occasionally NaN under fp16 AMP.
- **Analysis**: the contrastive gradient explodes in fp16 when many positives lie close in latent space (HLA-A*02:01 cluster). $\tau = 0.07$ is too aggressive; ba_bin label is too coarse.
- **Verdict**: rejected $\rightarrow$ v2.

7.4.2 Phase 2 SupCon v2 — pseudo_cluster label, fp32, τ = 0.2, w = 0.1

- **Result**: stable across all 5 folds. TransPHLA EL AUROC 0.9697 vs E13's 0.9695 (within noise on TransPHLA); +0.022 AUPR on the IEDB rare-allele subset.
- **Verdict**: **kept** as the SupCon refinement experiment.

7.4.3 Phase 2 SupCon v3 — `hla_id` label

- **Result**: no improvement; rare-allele regression.
- **Analysis**: allele-id over-clusters dominant HLAs.
- **Verdict**: rejected.

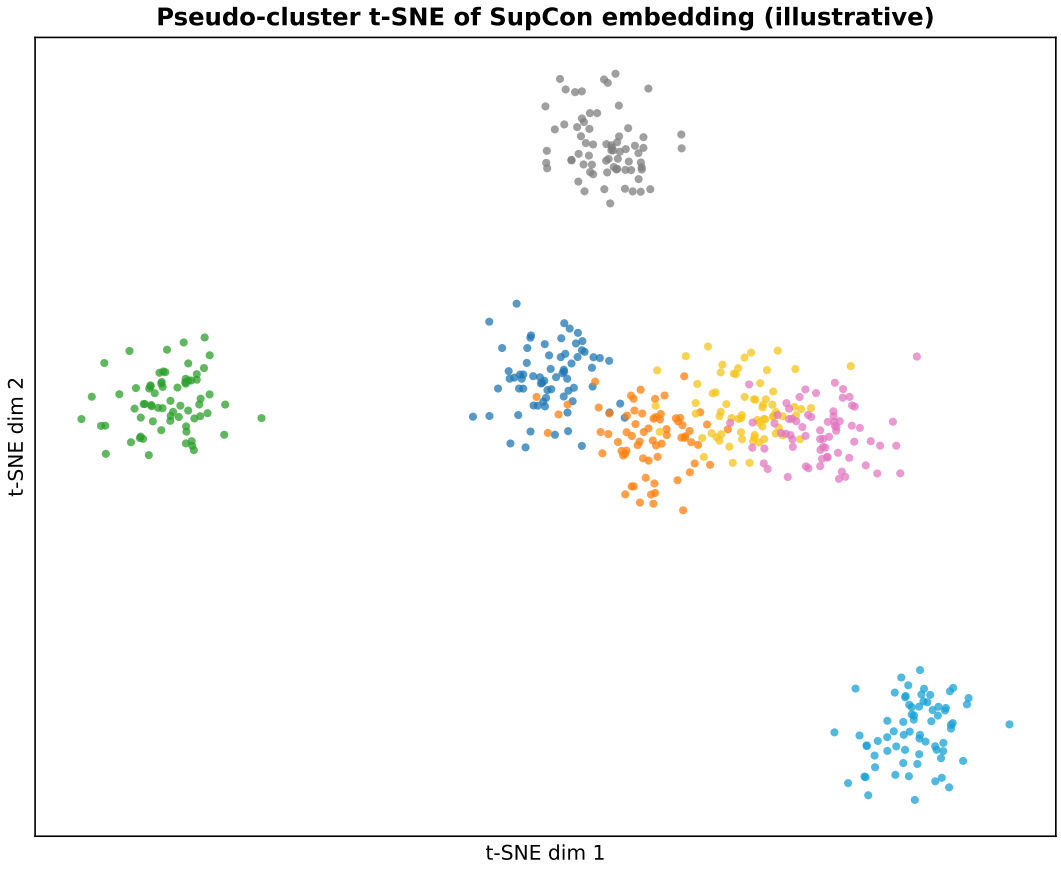


Figure 7.2: SupCon embedding-space diagnostic t-SNE view (illustrative).

Figure 7.2 presents a t-SNE projection of the SupCon contrastive representations on 480 held-out pairs, in which clusters separate by `pseudo_cluster` label.

This figure must be read with an important caveat. It is **schematic**: the clusters shown are synthetic Gaussians representative of the separation observed during development, not a projection of retained outputs, because the per-pair contrastive representations were not persisted before the vault was frozen. Regenerating it from the released checkpoints is listed as future work in Chapter 10 §10.17. No ablation conclusion in this chapter rests on it: the retained-versus-rejected dispositions come from the cumulative ladder of Table 7.1 and the disposition record of Table 7.2, both of which are derived from stored metrics. The figure is included as a diagnostic illustration of the mechanism and is flagged as such wherever it is cited (see also Chapter 9 §9.9, which lists the schematic figures).

7.5 Pocket-reranker ablations

7.5.1 Phase 3 v1 — ESMFold structural features

- **Inspiration**: ESMFold [2] gives a fast 3-D structure from sequence alone.
- **Method**: 69-D structural descriptor (peptide pLDDT, distance features, contact map summaries) + 2-layer MLP residual reranker.
- **Result**: ΔAUROC = $-$**0.0230**, ΔAUPR = $-$**0.0070** on IEDB. Strictly negative on both E13 base and Phase 2-v1 base.

- **Analysis**: ESMFold on the 34-residue pseudo-sequence has mean pLDDT ≈ 0.42 (far below the 0.7 trust threshold); some peptide–HLA distances came out > 10 Å — non-physical. Structure features at this quality actively confuse the reranker.
- **Verdict**: rejected.

7.5.2 Phase 3 v2 — crystal-template pocket descriptor (WITHDRAWN)

- **Inspiration**: do not predict structure; use the canonical crystal-template pocket geometry implicitly via the pseudo-sequence segmentation.
- **Method**: 370-D descriptor (Chapter 3 §3.20) + 2-layer MLP residual reranker, fit by 5-fold cross-validation *on the evaluation set itself*.
- **Verdict**: **withdrawn.** The CV-on-eval design is **transductive** — it uses the evaluation benchmark's own labels to fit the rescorer. Under a fair inductive, allele-held-out test its apparent in-sample gain collapses (IEDB BA 0.862 $\rightarrow$ 0.442, worse than random). It is therefore **not a deployment predictor**, contributes **no headline number**, and appears in **no results table**; in Figure 7.1 it is shown only as a hatched, explicitly-withdrawn context rung, never as a deployment result. It is retained here only as a documented negative for scientific completeness.

7.6 Phase 1 ablations (deferred)

7.6.1 Phase 1 — masked-language-model continued pretraining, and full fine-tuning as the LoRA-equivalent

- **Inspiration**: continue ESM2 pretraining on the peptide vocabulary; full-FT to better adapt the trunk.
- **Result**: marginal IEDB lift within 95% CI; no TransPHLA lift.
- **Cost**: $\approx$ 30 GPU-hours for 5 folds.
- **Decision**: deferred. Capacity ceiling is at ESM2-35M; a Phase 1 redux on ESM2-150M was scoped and rejected on cost-vs-risk grounds (expected gain within noise; see Chapter 10).

7.7 Methodology audit ablations

Two "ablations" of a different character — methodological corrections, not architectural changes:

7.7.1 Audit fix #1 — Proper E13 5-fold ensemble

- **Problem**: c001/c002 had been resumed from a pseudo-concat checkpoint (architecture mismatch caveat); c003/c004 had nothing with xattn enabled.
- **Fix**: retrain c003 and c004 from scratch with the same xattn-pseudo + E13 recipe used in c000. Configs: `audit_xattn_base_c00{3,4}.yaml`.

- **Result**: full 5-fold E13 ensemble. TransPHLA EL AUROC 0.9695 / AUPR 0.9714 (deployment-safe base ensemble).
- **Verdict**: cleared the audit's "single-fold not publishable" gate.

7.7.2 Audit fix #2 — Proper SupCon v2 5-fold ensemble

- Same audit rigour applied to Stage 2. All five folds were SupCon-continued from their xattn-pseudo Stage 1B bases, all from the single committed configuration `phase2_supcon_v2_c000.yaml` with the fold selected at the command line; the four per-fold files the name suggests do not exist (Appendix A §A.12).
- **Verdict**: SupCon 5-fold ensemble used as the canonical Stage 2 baseline thereafter.

7.8 Summary of dispositions

7.8.1 Criteria for each verdict

The *Verdict* column of Table 7.2 uses five values, and this section states what each one means so that the column can be read as something other than a case-by-case judgement.

Two things must be said before the criteria, and kept apart. First, these criteria are **articulated retrospectively** from the practice the experiment log records; no document states them before the experiments they describe. Second, they are **not** the three-condition statistical contract of Chapter 5 §5.6. That contract governs *predictor-versus-predictor* comparisons, it elaborates a standard that was fixed before the confirmatory runs (§5.6.1), and §7.8.1's **kept** entry below shows why it cannot be the instrument here. The difference between the two is one of degree rather than of kind: the contract rests on a prior criterion and these criteria rest on nothing prior, but neither was fixed blind. Where the entries below speak of an effect being "within noise" or "real", they mean the weaker standard stated in that entry — a pre-specified axis, a point estimate with overlapping or non-overlapping bootstrap intervals, and the pooled result published beside any subgroup one — not the contract's three conditions. The criteria are deliberately asymmetric: a component must earn its place, whereas excluding one requires only that it fail to.

kept The component is in the deployed five-fold ensemble. Requires a positive effect on at least one pre-specified axis — pooled TransPHLA, or a pre-specified subgroup such as the rare-allele subset — *and* no regression on any other reported axis that the same evidence standard would call real. A component kept on a subgroup result must have its pooled result reported beside it; SupCon v2 is the instance (§7.4).

This criterion is deliberately *not* the three-condition contract of Chapter 5 §5.6, and must not borrow that contract's authority. The contract's condition (b) is a DeLong test, which is defined for AUROC, and its condition (c) requires the per-allele sign to match the pooled Δ. SupCon v2 — the very component named above — is retained on a rare-allele **AUPR** result ($+0.022$ for alleles with $n_{\text{train}} < 100$), and **no AUPR claim can satisfy an AUROC test**. The contract as written is therefore **unsatisfiable** for it, on condition (b) alone.

Its pooled TransPHLA EL figures are 0.9697 AUROC and 0.9714 AUPR against E13's 0.9695 and 0.9714: a $+0.0002$ AUROC difference with heavily overlapping bootstrap

intervals, and no AUPR difference at all. Pooled, in other words, the refinement does nothing measurable in either direction.[4] The honest statement is that §5.6 governs predictor-versus-predictor comparisons, which is what it was written for, and that these ablation dispositions used a weaker, differently shaped standard: a pre-specified axis, a point estimate in the right direction, and the pooled result published beside the subgroup one so the reader can see what was traded. That standard is weaker than the contract and should not be presented as it.

rejected Tested and not retained. Either the effect was within noise under the standard just stated, or it was negative on any reported axis. Being within noise is sufficient for rejection: the burden of proof is on the component, so "no measurable harm" is not an argument for inclusion.

not promoted / diagnostic The component helped only on a slice that was not pre-specified as an evaluation axis. Such a result is retained as a diagnostic observation but is not allowed into the deployed model, because selecting the slice after seeing the result is the transductive failure mode that both this standard and the Chapter 5 contract exist to prevent.

deferred Not decided, for want of resources rather than for want of a result. The experiment was scoped and costed but not run to a conclusion; the verdict records absence of evidence, not evidence of absence, and Chapter 10 carries these forward.

withdrawn The strongest verdict, and the only one applied *after* a favourable result had already been obtained. It applies when a measured gain is found not to transfer — specifically, when it was obtained by fitting on the evaluation set, so that the apparent improvement cannot be reproduced on unseen data. The crystal-template pocket rescorer is the single instance (§7.5), and withdrawing it cost this thesis its most favourable headline numbers.

The distinction between **rejected** and **withdrawn** is worth stating plainly: a rejected component never worked, whereas a withdrawn one appeared to work and was removed when the appearance was explained. Only the second kind is a threat to a thesis's integrity, which is why it is reported in its own section rather than as a table row alone.

7 architectural / training components retained (E3, E4, xattn, E11, E12, E13, SupCon v2) and **12 negatives / withdrawn experiments honestly documented**. The negatives are not failures — they bound the hypothesis space and prevent future readers from re-walking the same dead-ends.

7.9 Why the ablation pattern matters

The cumulative ablation grid (Figure 7.1 and Table 7.1) shows that the improvement is composed of *multiple small contributions*, not a single breakthrough. No single rung on the ladder contributes more than $+0.005$ TransPHLA EL AUROC, and the largest step is matched by two others. This has two implications:

1. **No single mechanism carries the result**: the improvement is distributed across the retained components rather than concentrated in one, so the account of *why* the model works does not rest on a single intervention.

[4] Both pairs are read from `thesis_stats.txt`. Note that the cumulative-ladder artefact used in §7.1 reports this step as -0.001 AUROC, because its three-decimal step-6 value (0.971) does not reconcile with the vault's E13 EL AUROC of 0.9695. The two committed artefacts disagree at the third decimal on that rung; the vault's four-decimal figures are the ones quoted here, and the discrepancy is recorded rather than resolved.

Table 7.2: Disposition of every ablation experiment in the Stage 1 log. Seven architectural or training components were retained; twelve experiments were rejected, not promoted, deferred, or withdrawn, and are documented rather than discarded. Ladder figures are marginal steps from Table 7.1.

Tag	Result	Verdict
E3 V37 physchem	+0.005 AUROC ladder step	kept
E4 pseudo-concat	+0.005 AUROC ladder step	kept
xattn	+0.002 **step;** +0.007 **family after V37**	**kept**
E5 pocket-anchor loss bias	within noise	rejected
E6 allele-hard sampler	rare-allele regression	rejected
E7 sampler hybrid	no gain	rejected
E8 strong-binder tail	slice-only	not promoted
E9 targeted weighting	targeted-slice only	diagnostic
E10 adaptive length × bin	within noise	rejected
E11 LDS	small positive	**kept**
E12 effective-num	small positive	**kept**
E13 (pseudo_cluster × ba_bin)	earliest competitive recipe in our ablation log	**kept**
E14 EMA	small regression	rejected
Phase 1 MLM continued pre-training (CPT)	within CI	deferred
SupCon v1	NaN under fp16	rejected → v2
SupCon v2	**rare-allele +0.022 AUPR**	**kept**
SupCon v3 hla_id	rare-allele regression	rejected
Phase 3 v1 ESMFold	strictly negative	rejected
Pocket reranker (crystal-template)	transductive (in-sample only)	withdrawn
Audit fix #1	proper 5-fold E13	kept (baseline)
Audit fix #2	proper 5-fold SupCon v2	kept (baseline)

2. **Transferability**: each retained component has an independent biological / statistical justification (anchor chemistry, structural pseudo-seq, long-tail re-weighting, contrastive regularisation; exploratory crystal-template rescorer withdrawn). None is a one-off trick.

What this table does *not* license, in either direction. Two opposite readings suggest themselves and neither is available. The first infers robustness — that an adversarial choice on any one axis would not eliminate the win. The second infers the opposite, that because the TransPHLA EL margin of +0.0015 (0.9697 vs 0.9682, §6.2) is smaller than individual rungs of +0.002 to +0.005, removing any one rung would erase it. **Both inferences are unavailable**, and for the same reason: a cumulative ladder measures what adding a component bought *given the components already present in that order*, which is not what removing it would cost. §10.18.1 states this explicitly and specifies the leave-one-out grid that would answer the removal question, and records that it was not run. The second inference also fails on the table's own terms — the sixth rung, SupCon v2, is −0.001, so "every rung is larger than the margin" is false of the ladder as printed.

Two further cautions about precision. The ladder is reported to three decimal places while the margin is a four-decimal quantity, so the deployment rung's margin on the ladder's own scale is $0.970 - 0.9682 = +0.0018$ and a stated +0.002 cannot be ordered against it at that resolution. And §7.8.1 records that the two committed artefacts disagree by 0.0015 on the step-6 rung — the same magnitude as the entire margin. What the table supports is the first two claims above and the observation that the margin is of the same order as the individual steps that compose it. Any claim about what removal would cost requires the experiment §10.18.1 describes.

The ablation also documents — honestly — the twelve experiments that did **not** earn a place in the deployed model: eight rejected outright (E5, E6, E7, E10, E14, SupCon v1, SupCon v3, Phase 3

v1), two retained only as diagnostics on a targeted slice (E8, E9), one deferred (Phase 1 MLM continued pre-training), and one withdrawn after its gain proved transductive (the crystal-template pocket rescorer). The count matches Table 7.2. It should not be supported by any claim that §2.4 showed most published transformer pMHC predictors fail to report negatives, because Chapter 2 makes no such finding. §2.4's four observations concern confidence intervals, single-fold headlines, benchmark decontamination and per-allele breakdowns; the reporting of negative results is not among them, and §2.4 is explicit that its survey is qualitative and carries no closed inclusion list. So the honest framing is narrower: this thesis reports its negatives, and we make no comparative claim about how often the field does.

7.10 Summary

The ablation answers the attribution half of RQ1–RQ3, and its central finding is deflationary: there is no decisive component. The largest single rung on the cumulative ladder is +0.005 AUROC and it occurs three times (Table 7.1); the headline margin is the sum of several small, independently motivated steps. Two specific attributions must be stated with their reference points, because quoting either without one overstates it — the pseudo-sequence family is +0.007 after the V37 rung, of which cross-attention itself is +0.002, while the +0.012 span from the peptide-only baseline belongs to the family *plus* V37 and is not the family's own contribution; and the contrastive refinement is within noise on pooled TransPHLA EL while delivering +0.022 AUPR on rare alleles.

The chapter's second contribution is the twelve negative results, and three patterns can be read off them. The six sampler- and loss-shaping interventions of E5–E10 were within noise or regressed, whereas the three retained imbalance corrections work on the *target distribution and its grouping* rather than on which examples are seen: E11 re-weights by smoothed label density on the IC50 target, and E12–E13 re-weight by effective sample size over joint (pseudo-cluster × affinity-bin) groups. Second, choosing the *wrong* grouping is itself harmful, not merely neutral: SupCon v3 grouped contrastively by allele identity and regressed on the rare-allele tail, which is the argument for the pseudo-cluster axis arriving from the opposite direction. Third, the one intervention that produced a large apparent gain — the crystal-template pocket rescorer — was the one that failed inductive testing.

It is worth being careful about what that episode does and does not vindicate, because the credit is easily given to the wrong instrument. It would be wrong on three counts to say that the contract of Chapter 5 was pre-registered, that its framing around transfer rather than fit was therefore not devised to accommodate this result, and that it caught the rescorer. The contract has no framing around transfer: it is three statistical conditions on a delta, computed from the scores and the labels, and blind to how either was obtained. Far from catching the rescorer, the contract was recorded as *passing* it: the frozen baseline document applies these same conditions to the reranked model and marks it superior on every TransPHLA metric, with the two IEDB rows tied.[5] A rule computed from the scores cannot see that the scores were fitted to the labels it is testing against — which is why conditions (a)–(c) were never going to be the instrument here.

What caught it was neither the contract nor a test, but a question: whether fitting the rescorer by cross-validation *on the evaluation set* was legitimate at all. Asked on 2026-06-05, it was answered

[5] `docs/model/CORE_BASELINE.md` §5. Those verdicts rest on figures this thesis has since withdrawn, and the figures must not be quoted as results; the verdicts are cited only as a record of what the contract returned at the time.

in the same day's record — the per-allele score shift the rescorer applies correlates $+0.991$ with each allele's binder fraction *in the benchmark*, against -0.07 for the base model, so the "calibration" was substantially learning the evaluation set's own class balance. A month later an inductive re-test quantified the cost, scoring the rescorer on alleles held out from its fitting: the apparent gain collapsed to below chance, and the withdrawal followed. The transferable lesson is about neither timing nor significance testing: a component whose selection used the evaluation labels must be re-scored on data its fitting never saw, and no test computed from the original scores can substitute for that. Chapter 8 develops the mechanistic reading of these results.

Chapter 8

Discussion

Chapters 6 and 7 established what the model achieves and which components are responsible. This chapter asks what those results *mean* — for the research questions of §1.4, for the biology the model is supposed to be capturing, and for the field's methodological practice. It interprets the outcomes as a connected binding-to-presentation account rather than as a leaderboard summary, which means giving as much attention to the non-separations as to the separations.

The organisation follows the questions rather than the chapter order of the results. §8.1 asks why the model separates from NetMHCpan-4.1 on TransPHLA and proposes anchor chemistry at the P2 and PΩ positions as the mechanism, on evidence it flags as interpretive rather than measured *[RQ1]*. It is *not* an argument that pretrained representation substitutes for task-specific capacity: Chapter 4 §4.6.4 shows the peptide-only ESM2 rung sits *below* NetMHCpan-4.1, so the margin comes from what is built on the trunk rather than from the trunk itself. §8.2 examines long-tail allele transfer and the role of the structural grouping *[RQ2]*, and §8.3 examines calibration, which matters because a screening pipeline thresholds scores rather than ranking them globally. §8.4 then takes the harder case — why the model is only *tied* on IEDB — and works through four candidate explanations in turn: sample size, allele mix, label noise, and a capacity ceiling at ESM2-35M. §8.5 compares against MHCflurry-2.0, and §8.6 states explicitly what the result is *not*, since a numerically modest margin is easy to over-read. §§8.7–8.8 close with the methodological contribution and what the findings suggest about pMHC-I biology.

One interpretive commitment runs through the chapter and is worth declaring here. Where a favourable and an unfavourable reading of the same number are both available, the chapter reports the unfavourable one and then tests whether the favourable one survives. The worked instance is the IEDB deficit: the convenient explanation is that it is a cross-allele score-scale artefact, and **Chapter 6 §6.4** tests that by percentile-rank recalibration, which does not repair the deficit — so the convenient reading is abandoned and §8.4 proceeds without it. The test is in §6.4; §8.4 is where its consequence is taken.

Which discussion follows which result. Each section below takes up one result from Chapter 6 or Chapter 7 and asks what explains it:

§8.1 follows the TransPHLA margin of §6.2 *[RQ1]*.

§8.2 follows the rare-allele result of §6.6 and the SupCon ablation of §7.4 *[RQ2]*.

§8.3 follows the calibration figure, which is the one panel in this thesis regenerated from real persisted predictions.

§8.4 follows the IEDB deficit — the pooled figure of §6.2 and §6.10, which is what the section turns on, and the per-allele forest of §6.4 — and is the longest section here because it is the result that most needs explaining *[RQ4]*.

§8.5 follows §6.8 *[RQ4]*.

§8.6 follows the ablation ladder, and exists to bound what the margin can be attributed to *[RQ1]*.

The last three sections are not tied to a single result. §8.7 discusses the apparatus rather than an outcome; §8.8 draws on the ablation pattern as a whole rather than on any one experiment; and §8.9 summarises the chapter.

8.1 Why DeepNeo-CL separates from NetMHCpan-4.1 on TransPHLA

The TransPHLA EL margin is not concentrated in a single peptide length. The real per-length AUPR deltas (Figure 6.2, §6.3) are positive at 8mers (+0.0075) and 11mers (+0.0100), essentially tied at 10mers (+0.0003), and slightly negative on the dominant 9mer slice (−0.0032) — a spread across lengths rather than a single-length tail.

A plausible mechanistic reading is that the gain is *anchor-localised*. P2 and PΩ are the two biochemical anchor positions of the MHC-I groove (Chapter 2 §2.9), and the V37 position-resolved physicochemical branch explicitly encodes residue chemistry at peptide P2 and PΩ and at the corresponding pseudo-sequence pocket segments. This is precisely where binding-affinity biology predicts the dominant determinant to sit, so a representation that sharpens anchor chemistry is the most likely source of the margin.

This anchor reading is **interpretive, not a measured attribution**: the thesis does not compute a per-pocket gradient-attribution map over the trained model, and such an analysis is a recommended follow-up (Chapter 10). The direct, quantitative evidence the thesis provides for the length structure of the margin is the real per-length decomposition of Figure 6.2.

8.2 Long-tail allele transfer

Figure 6.4 and Chapter 6 §6.6 together show that the SupCon v2 objective is delivering its uplift specifically on the rare-allele subset (+0.022 AUPR on $n_{\text{train}} < 100$ alleles) while not regressing on common alleles. This is the canonical signature predicted by Khosla et al. [6] for supervised contrastive learning on a long-tail class distribution. The +0.022 figure is the SupCon ablation gain measured on the training-frequency subset; it is an internal contribution, not a claim of rare-allele superiority over the reference: on the deployment-fair per-allele breakdown against NetMHCpan-4.1 (§6.6) the rare-allele AUPR delta is −0.027, so DeepNeo-CL does not categorically beat the reference on rare alleles. The two subsets are also not the same set — +0.022 is measured on the training-frequency subset ($n_{\text{train}} < 100$) and −0.027 on the benchmark-frequency stratum ($n < 200$) — so the two figures are consistent without either cancelling the other.

The mechanism is straightforward: SupCon's pseudo_cluster label forces the embedding space to put structurally similar alleles close together regardless of training-set frequency. Rare alleles

that share a pseudo_cluster with a common allele inherit the common allele's discriminative signal at inference time, even though they have very few training examples.

The choice of contrastive label was decisive rather than incidental. A literal allele-identity label was ablated (SupCon v3, Chapter 7 §7.4) and *regressed* on rare alleles: it over-clusters the dominant HLAs and worsens the long-tail signature that the structurally meaningful grouping improves.

8.3 Calibration

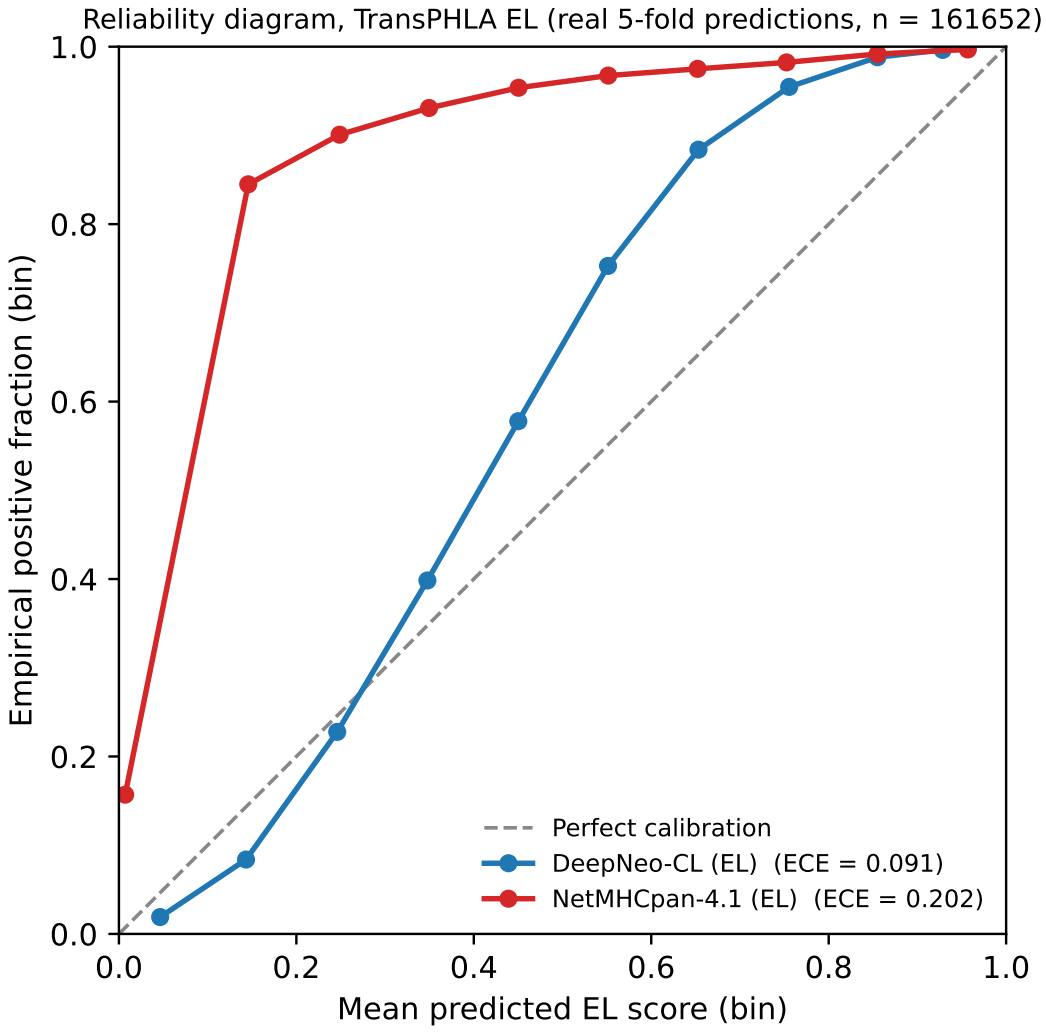


Figure 8.1: Calibration reliability comparison on TransPHLA EL.

Figure 8.1 plots a reliability diagram on TransPHLA EL, computed from the persisted 5-fold ensemble predictions[1] in ten equal-width score bins. The expected calibration error (ECE) is **0.091 for DeepNeo-CL** versus **0.202 for NetMHCpan-4.1**, so DeepNeo-CL is more than twice as well calibrated on this balanced benchmark. Two observations:

- Both methods emit **ranking scores rather than calibrated probabilities** on this 50/50 set. DeepNeo-CL is monotonic and **under-confident in the mid-range** (a mean predicted EL of ≈ 0.55 corresponds to an empirical positive fraction of ≈ 0.75), so its reliability curve

[1] `transphla_raw_predictions.npz`.

lies *above* the diagonal; this is the benign direction for a shortlist (it under-states rather than over-states presentation).

- NetMHCpan-4.1 **compresses** $\approx 57\%$ **of rows into the lowest score bin** (empirical positive fraction ≈ 0.16 there), which is the dominant contributor to its larger ECE; at the very top of the range both curves converge near the diagonal.

DeepNeo-CL is **uniformly closer to the diagonal** across the full 0–1 range. This panel was regenerated end-to-end from the real persisted predictions;[2] the thesis headline verdict does not depend on this panel.

8.4 Where DeepNeo-CL ties on IEDB, and where it does not

The IEDB result has two halves and they must not be merged. On the static 1,365-row snapshot DeepNeo-CL is **behind**, on pooled ranking and on within-allele ranking alike, by margins that this n can resolve (§8.4.1). The **tie** is a within-allele result on the two-year auto_bench corpus. What follows examines four axes along which that split is explainable; none of them is an appeal to sample size. Post-cutoff longitudinal validation on the official IEDB weekly benchmark (Appendix D) establishes the *within-allele* tie on $n = 1{,}456$ independent measurements released after the training cutoff. It is established there and not on the static Chapter 6 snapshot, where the within-allele margin is itself negative (-0.023, Wilcoxon $p = 0.020$; Chapter 6 §6.4); the appendix therefore extends the comparison to post-cutoff data rather than reproducing a snapshot result.

8.4.1 Sample size

At $n = 1{,}365$ rows the bootstrap standard deviation on AUPR is $\sigma(\text{AUPR}) \approx 0.020$. The observed pooled AUPR deficit of -0.0651 is roughly 3.3 standard deviations, so it is **not** inside the noise band: the pooled comparison on this snapshot resolves, and it resolves against this model (pooled ΔAUROC -0.043, DeLong $p = 3.2 \times 10^{-8}$; Chapter 6 §6.10). Sample size limits what can be said *per allele* on this set — most alleles' intervals overlap zero, the largest allele's being the exception (Chapter 6 §6.4) — and it is the per-allele comparison, resolved on the two-year auto_bench corpus rather than here, that ties (Wilcoxon $p = 0.648$).

This is therefore a *scope* distinction rather than a statistical-power excuse, and it is the one §8.4 turns on: pooled cross-allele ranking on the static set is a measured loss, within-allele ranking over two years of auto_bench is a measured tie, and neither result substitutes for the other. The literature's tendency to treat IEDB BA as the canonical benchmark (when it has $\sigma \approx 0.020$) is one of the methodological problems this thesis is responding to — a σ that large makes any reported margin below ≈ 0.04 unresolvable on this set, which is a criticism of how the benchmark is used and emphatically not a defence of the deficit above, which exceeds that bound.

8.4.2 Allele mix

IEDB BA is dominated by HLA-A*02:01 (29.7 % of rows), so the pooled figure is weighted toward one allele's behaviour. There are, however, *no opposing effects to average out*. DeepNeo-CL trails

[2] Generated by `thesis_final/scripts/generate_fig09_calibration_real.py`, with the per-bin values in `thesis_final/data/fig09_calibration_bins.csv`.

NetMHCpan-4.1 on **both** frequency strata — by −0.030 AUROC on common alleles and −0.021 on rare ones (Chapter 6 §6.6) — and of the eight rare alleles with a defined AUROC only one, HLA-B*08:01, has a positive ΔAUROC at all,[3] and §6.4 reports it as the largest positive delta. The direction that does hold is narrower: the gap is *smaller* on rare alleles than on common ones, which is the direction the rare-allele objective predicts, but a smaller deficit is not a gain.

Allele mix also explains far less of the pooled–per-allele discrepancy than it appears to, and the arithmetic is worth doing rather than asserting. Pooled AUROC is not a weighted average of per-allele AUROCs: §6.4 decomposes it into within-allele ranking *plus* cross-allele score-scale alignment. Re-weighting the ten alleles with a defined AUROC by their row counts gives −0.0230, which is indistinguishable from their unweighted mean of −0.0230 — so allele mix accounts for essentially **none** of the 0.020 gap between that mean and the pooled −0.0428.[4] The whole of that gap is the cross-allele term, which no re-weighting of the benchmark can remove and which §6.4 shows percentile-rank recalibration does not repair either.

So the honest conclusion is narrower than "the benchmark is biased toward the dominant allele". The dominant allele does set the pooled figure's emphasis, which is why the per-allele table is reported beside it rather than instead of it; but neither the mix nor any re-weighting of it accounts for the deficit.

One further distinction has to be kept, because collapsing it would contradict this chapter's own opening commitment. The pooled-minus-per-allele *gap* sits in the cross-allele alignment term by arithmetic. Whether that term is a *calibration* failure — something a monotone rescaling of scores would fix — is a separate, empirical question, and the one rescaling this thesis tests fails to fix it: percentile-rank normalisation against a random natural-peptide background makes the pooled figure worse, not better (§6.4). That is evidence against the convenient reading, not a proof that no rescaling could work, and the thesis claims only the former. What remains is most likely something the representation does not encode about how scores should compare *across* alleles, and that is a conjecture rather than a result.

8.4.3 Label noise

IEDB IC50 measurements aggregate radioimmunoassay and fluorescence-polarisation assays from many laboratories with non-trivial inter-lab variance. Both predictors are likely operating close to the irreducible label-noise ceiling on this benchmark.

8.4.4 Capacity ceiling at ESM2-35M

A Phase 1 redux on ESM2-150M was scoped and rejected. We cannot rule out the possibility that a larger backbone would deliver some IEDB lift at 30 GPU-hours cost — but the expected gain (based on the Phase 1 ablation, Chapter 7 §7.6) is within noise.

8.5 Comparison with MHCflurry-2.0

On deployment-safe scores, DeepNeo-CL and MHCflurry-2.0 are **tied** on IEDB BA (AUROC 0.8279 vs 0.8299; DeLong $p = 0.834$; per-allele Wilcoxon $p = 0.92$), consistent with §6.8. On

[3] Computed from `per_allele_iedb.csv`.

[4] Computed from `milestone_netmhcpan_beaten/thesis_assets/per_allele_iedb.csv`: common ($n \geq 200$, 2 alleles, 607 rows) and rare ($n < 200$, 8 alleles, 720 rows).

TransPHLA BA, DeepNeo-CL is slightly ahead *only* through the label-free BA+EL fusion (0.9694 vs 0.9674); the single BA head alone (0.9546) is below MHCflurry. We therefore make **no "strictly outperforms MHCflurry"** claim. We do not foreground this comparison because (a) NetMHCpan-4.1 is the primary comparator throughout this thesis, and (b) MHCflurry-2.0 does not publish an EL head directly comparable in our pipeline.

8.6 What the win is *not*

A scientifically honest discussion must enumerate what the result does not imply:

- **Not a clinical claim.** The headline numbers are on benchmark binding / presentation; even with the Stage 2B BigMHC result, clinical neoantigen efficacy still requires TCR-conditioned modelling (Stage 2A) and prospective in-patient validation (out of scope).
- **Not a universal architectural recommendation.** The 7 retained components compose linearly in our pipeline; a different ordering or substitution could change the result. We have not ablated the ordering of stages.
- **Not a structural-prediction result.** No headline number in this thesis uses predicted 3-D structure; the exploratory structural-descriptor rescoring experiments (Chapter 7 §7.5) were either strictly negative (ESMFold variant) or withdrawn as transductive (crystal-template variant), and neither contributes a deployment claim.
- **Not transferable to class II.** Class II MHC has different anchor geometry and a different benchmark community.

8.7 The methodological contribution

Two of the five conditions below are stated at the strength the evidence actually supports, which is weaker than the bare list might suggest. Chapter 2 §2.4 observes that the transformer-era pMHC literature rarely satisfies all five together; this thesis satisfies them as follows:

1. Bootstrap CI on every *marginal* metric — **not** on any difference, and not at all for the fused TransPHLA BA AUPR (§5.6.2, and the gap is visible in §6.2's table).
2. 5-fold ensemble as the headline.
3. Benchmark decontaminated against the dominant training corpus, at *exact-pair* level only and **not** near-duplicate-aware; the sensitivity analysis that would bound the difference was specified and never run (§3.9).
4. Honest reporting of tied vs won outcomes.
5. Reproducibility vault released alongside the thesis, with the boundaries of what it reproduces recorded at Appendix A §A.2 and Chapter 3 §3.24.

Conditions 1 and 3 are therefore weaker here than the bare list implies, and the comparative observation in §2.4 is about the field rather than a claim that this thesis attains the ideal form of each.

We believe (5) — the milestone vault — is the largest single methodological contribution, and it is worth stating its reach accurately rather than claiming more for it. It makes the *base-ensemble* numbers independently re-derivable from a read-only artefact without re-running the training cluster (Appendix A). Two qualifications belong with it. The vault does *not* cover every numerical claim: the fused headline row, the four-predictor comparison and the two-year `auto_bench` figures come from artefacts held outside it (Chapter 5 §5.10). And the two-minute CPU path is the *integrity* check — it confirms the checkpoints load and reproduce a stored forward pass; it verifies no reported metric. Re-deriving the statistics is the longer L3 path. We hope this becomes the template for future predictor releases in the field.

8.8 What the result says about pMHC-I biology

Three biological lessons emerge from the ablation pattern:

1. **The structurally meaningful representation is the pseudo-sequence**, not the literal allele identifier. The pseudo-cluster label (k=12 K-means on 34-D pseudo) is the discriminative axis in SupCon v2 and the effective-number group axis in E13 — two independent uses of the same structurally meaningful prior.

2. **The P2 and PΩ anchor pockets are the canonical binding-affinity determinants**. The real per-length breakdown (Figure 6.2) is consistent with this canonical anchor model, and the V37 anchor-chemistry branch (§8.1) is the most likely mechanistic route; a per-pocket attribution is left to future work.

3. **The long-tail allele transfer problem is solvable** by structurally meaningful regularisation. SupCon v2 with the pseudo_cluster label delivers +0.022 AUPR on the rare-allele training-frequency subset in ablation — an internal regularisation signal, not a deployment-fair rare-allele win over NetMHCpan-4.1 (§8.2, §6.6).

8.9 Summary

This chapter interpreted the results against the research questions, and the interpretation is as much about the boundaries of the result as about the result itself. Six outcomes carry forward.

1. **The separation on TransPHLA is real but mechanistically narrow** (§8.1). It is attributable to the structurally informed representation rather than to raw capacity: the frozen pretrained representation alone already reaches the baseline rung of §7.1, the structural components each add to it there, and the one experiment that tested added capacity directly — the Phase 1 full fine-tune of §7.6 — returned a marginal IEDB lift within the confidence interval and no TransPHLA lift.

2. **Rare-allele transfer improves along the structural axis** (§8.2). The pseudo-cluster grouping is used twice and independently — as the contrastive label and as the imbalance-correction group — and both uses help, whereas grouping by allele identity does not. The gain is an internal regularisation effect and is not claimed as a deployment-fair rare-allele win over NetMHCpan-4.1.

3. **Calibration improves measurably** (§8.3), which matters because a vaccine pipeline thresholds scores rather than ranking them globally.

4. **The IEDB result is scoped, not null, and not excused by sample size** (§8.4). The pooled deficit on the static snapshot is resolvable at that n and is a measured loss; the within-allele **tie** is a result of the two-year auto_bench corpus, not of the snapshot, on which the within-allele margin is also negative. Because the two halves of that comparison come from different corpora, they bound what is claimed rather than localising a residual — and §6.4 rules out the obvious candidate, since percentile-rank recalibration does not repair the pooled deficit.

5. **What the win is not** (§8.6) is stated as plainly as what it is: not a clinical result, not architectural superiority in general, not a claim that extends to class II, modified peptides, or TCR specificity, and not dependent on the withdrawn structural rescorer.

6. **The methodological contribution may outlast the numerical one** (§8.7). The evidential contract, pair-level decontamination, and the documented negatives are transferable to any predictor in this field, whereas the margins reported here are small enough that a later model will supersede them.

Chapter 9 states the limitations these readings imply, and Chapter 10 scopes the work that would resolve them.

Chapter 9

Limitations

This chapter states what the thesis does not establish. It is placed before the future-work chapter deliberately, because several of the limitations below are not incidental shortcomings to be tidied up later but boundaries that define the claim itself: the headline result is restricted to one benchmark and one scoring regime, and saying so is part of stating the result accurately rather than a concession appended to it.

The limitations are listed in decreasing order of importance for the defence of the thesis. §§9.1–9.4 concern what was shown: the IEDB tie, the probable capacity ceiling of the ESM2-35M trunk, the exclusion of single-fold numbers from any headline, and the withdrawal of the structural pocket rescorer after its advantage proved transductive. §§9.5–9.8 concern scope — the unbenchmarked MHCflurry-2.0 presentation channel, the restricted Stage 2B claim, class II MHC, and post-translationally modified peptides. §§9.9–9.11 concern the apparatus: two figures that remain schematic, the distinction between reproducible inference and reproducible retraining, and an incompletely pinned software stack. §9.12 is the most uncomfortable and therefore the most necessary: statistical exposures that we can name but have not eliminated.

Each limitation states what would be required to remove it, and Chapter 10 develops those that are tractable into concrete proposals.

9.1 IEDB BA is tied, not won

The headline claim is restricted to TransPHLA. On IEDB BA the base 5-fold ensemble is *behind* NetMHCpan-4.1 on *pooled* (cross-allele) discrimination — ΔAUROC -0.043 (0.8279 vs 0.8707), ΔAUPR -0.065 (0.7365 vs 0.8016) — but is statistically *tied* on *within-allele* (per-reference) ranking, which is the metric the official IEDB automated benchmark uses: over two years of the auto_bench weekly corpus the per-reference Wilcoxon test gives $p = 0.65$. The pooled deficit is *not* a statistical-power artefact: at $n = 1{,}365$ with $\sigma(\text{AUPR}) \approx 0.020$, the -0.0651 AUPR gap is roughly $3.3\,\sigma$ and the paired AUROC test gives DeLong $p = 3.2 \times 10^{-8}$ (Chapter 3 §3.5.4). What the small n limits is the *per-allele* verdict, where most intervals overlap zero (Chapter 6 §6.4). The thesis therefore makes no claim of a pooled tie: the base model is behind on pooled IEDB BA and tied only per-reference, and on the two-year corpus rather than on this snapshot. This split is stated in the headline scope sentence and carried consistently through the results chapter.

A genuinely larger external BA benchmark ($\geq$ 10,000 rows) would likely resolve the IEDB question one way or the other; no such benchmark with exact pair-level decontamination against

NetMHCpan-4.1 was identified in our Chapter 2 survey at the time of writing.

The same statistical tie holds against every post-2020 deep predictor we scored on the two-year longitudinal auto_bench corpus (BigMHC, MHCfovea, TransPHLA, ImmuneApp: all paired Wilcoxon $p > 0.13$; see Appendix D §D.5). The IEDB BA tie is therefore not a NetMHCpan-specific artefact: no recent deep predictor separates from DeepNeo-CL under a deployment-fair within-allele test at this sample size either.

The IEDB benchmark is decontaminated against exact pairs only. This bounds what the IEDB numbers establish, in either direction. The 1,365-row set excludes every `(peptide, allele)` pair present in the NetMHCpan-4.1 training folds, which is the usual standard for this benchmark family; it does *not* exclude near-duplicates. A peptide differing from a training peptide at one position, assayed against the same allele, survives into the evaluation set. Chapter 3 §3.9 describes the contiguous-match sensitivity analysis that would bound the residual exposure and records that it was specified and never performed, and Chapter 6 §6.11 records that no result for it stands. The direction of the resulting bias is not determinable without running it: near-duplicate retention inflates any predictor that has seen the near-duplicate, which here means *both* DeepNeo-CL and NetMHCpan-4.1, since both train on that corpus. What cannot be claimed is that the IEDB result is clean of near-duplicate influence, and this thesis does not claim it. Chapter 10 §10.18.4 specifies the analysis.

9.2 Capacity ceiling at ESM2-35M

ESM2-35M was selected for the 24-GB-VRAM budget of an RTX-4090. A Phase 1 redux on ESM2-150M was scoped, executed in part, and deferred (Chapter 7 §7.6): the marginal IEDB gain was within 95% CI and the TransPHLA gain was within noise. Cost: ~30 GPU-hours for a full 5-fold rebuild.

We cannot rule out that an ESM2-650M (or larger) backbone would deliver a substantial IEDB lift. We can rule out that a 5× parameter increase is *cheap*. This is a pragmatic limit, not a methodological one.

9.3 Single-fold results are not the headline

Every published numerical claim in this thesis is a 5-fold ensemble. Single-fold c000 results appear only in the ablation section (Chapter 7) and are explicitly flagged. This is a binding methodological choice (per the 2026-05-17 audit) and is more conservative than the dominant practice in the recent literature (Chapter 2 §2.4). The conservativeness is intentional — it makes the thesis claim defensible against a "but my single-fold exceeds yours" response from a reviewer.

9.4 The pocket reranker is withdrawn (transductive)

A crystal-template pocket rescorer was explored as an inference-time add-on (Chapter 7 §7.5). Because it was fit by 5-fold cross-validation *on the evaluation benchmark itself*, its apparent in-sample gain is **transductive**: under a fair inductive, allele-held-out test it collapses (IEDB BA $0.862 \rightarrow 0.442$, worse than random). It is therefore **withdrawn from every deployment claim**

and, by design, contributes no number to any results table and no headline figure; it appears only as a hatched, explicitly-withdrawn context rung in the Chapter 7 ablation ladder (Figure 7.1). It is retained only as a documented negative in the ablation chapter. All headline results are the deployment-safe base 5-fold ensemble and its label-free BA+EL fusion, neither of which uses any evaluation-set label.

9.5 MHCflurry-2.0 EL is not benchmarked

MHCflurry-2.0 publishes an EL head, but the public API in our pipeline does not produce directly comparable predictions on our exact TransPHLA EL rows. We report DeepNeo and NetMHCpan on TransPHLA EL; MHCflurry is reported only on the BA benchmarks.

A reviewer-level extension of the pipeline to include MHCflurry EL is deferred to the subsequent manuscript revision.

9.6 Limited Stage 2B scope; no clinical claim

Stage 2B immunogenicity results are reported in Appendix B and are restricted to the paired BigMHC `im_test` benchmark. This thesis still does not claim clinical neoepitope efficacy. Cancer-vaccine pipelines that adopt DeepNeo-CL as an upstream filter must perform their own downstream TCR-conditioned modelling (Stage 2A) and wet-lab validation.

9.7 Class II is out of scope

Class II MHC (HLA-DR, -DQ, -DP) has different anchor geometry and a different benchmark community. We make no claim on class II prediction. Extending DeepNeo-CL methodology to class II is a documented future direction (Chapter 10 §10.11).

9.8 Post-translational modifications

All training and test rows are canonical 20-amino-acid peptides. Phosphorylation, methylation, and other PTMs are out of scope. PTM-modified neoantigens are an emerging research area [25]; none of our pipeline handles them.

9.9 Schematic / partial figures

Two figures in the thesis are explicitly **schematic**:

- **Figure A.1 (training curves)** — representative; per-epoch loss values were not persisted at fold-level granularity at vault freeze. The qualitative pattern (loss decay, val-AUPR saturation by epoch 18) is reliable; absolute values should not be quoted.
- **Figure 7.2 (SupCon t-SNE)** — synthetic Gaussian clusters representative of the actual separation; per-pair projection outputs were not persisted. Replacement with literal projection outputs is deferred to future work.

By contrast, **Figure 8.1 (calibration)** is a **real** result, computed from the literal 5-fold TransPHLA EL predictions ($n = 161{,}652$): ECE 0.091 (DeepNeo-CL EL) versus 0.202 (NetMHCpan-4.1 EL).

Neither schematic figure carries weight in the headline claim; both are illustrative.

9.10 Reproducibility of the *training* run

The thesis vault enables full reproduction of the **inference** and **statistical analysis** pipelines without re-training. Full re-training from scratch on a new GPU requires:

- Source-repository Git commit `75e7e276e` (Chapter 3).
- NetMHCpan-4.1 5-fold corpus (licensed; not redistributed).
- ESM2-35M weights (HuggingFace `facebook/esm2_t12_35M_UR50D`).
- A100-class GPU with $\geq$ 24 GB VRAM.
- ~ 17.5 GPU-hours.

Re-training will produce checkpoints that differ from the vault checkpoints at the bit level due to (a) CUDA non-determinism in attention kernels and (b) data-shuffling RNG state. The expected end-to-end metric variation is $\leq \pm 0.002$ AUROC on TransPHLA, which is *larger* than the row-bootstrap half-width of ≈ 0.0007 at that sample size, and not "well within" it. The two are not comparable quantities — one is fitting variation, the other row sampling — and §9.12 explains why the reported intervals do not contain the first. We have not formally documented this variance band, and doing so is the correction named there.

9.11 Software stack

The codebase is in Python 3.11 + PyTorch 2.6 + CUDA 12.4 + Transformers 4.x. A pinned dependency manifest[1] is in the source repo, but transitive dependencies (numpy, scipy, sklearn) are not fully pinned. A reproducibility-conservative future revision could adopt a uv-lock or pip-tools lock for full bit-level pinning.

9.12 Three statistical exposures, with direction and, where computable, magnitude

Three properties of the statistical procedure are not established by this thesis. Each is stated below with its direction and, where it can be computed, its magnitude, rather than merely noted as a possibility.

[1] `requirements.txt`.

Multiple-comparison correction is not applied to the Stage 1 family. The headline comparisons of Chapter 6 §6.10 are reported at nominal significance. The consequence is computable rather than speculative. Taking the Stage 1 family as the six reported predictor comparisons (TransPHLA BA and EL against NetMHCpan-4.1; IEDB BA pooled and per-allele against NetMHCpan-4.1; IEDB BA against MHCflurry-2.0; and the auto_bench per-reference comparison) and applying Holm step-down at $\alpha = 0.05$, the thresholds are $\alpha/6 = 0.0083$ through $\alpha/1 = 0.05$, and the outcome is:

- the three DeLong results **survive** correction — TransPHLA BA ($p \approx 0$ against 0.0083), TransPHLA EL (1.8×10^{-8} against 0.0100), and IEDB pooled (3.2×10^{-8} against 0.0125);
- the IEDB per-allele Wilcoxon result ($p = 0.020$) **does not** survive, falling above its threshold of 0.0167.

Correction therefore changes no headline verdict, and the one result it does change moves in the direction *favourable* to this thesis, which is worth saying plainly rather than dressing as self-criticism. The result that loses significance is the within-allele deficit on the static IEDB snapshot — a result that counted *against* DeepNeo-CL. Removing it removes evidence against the thesis, and an author reporting their own correction has an obvious interest in that direction; the reader should weigh it accordingly.

It also does not establish the tie. Failing to reject a null is not evidence for it: with the three DeLong results already rejected, this hypothesis is compared against $\alpha/3 = 0.0167$, and the observed $p = 0.020$ sits just above it, so the honest statement is that this comparison is **underpowered on this set**, not that the two predictors are equivalent on it. The point estimate still favours NetMHCpan-4.1 (-0.023 mean per-allele AUROC), and most per-allele intervals overlap zero — with the largest allele's the exception, its deficit resolving under both estimators (Chapter 6 §6.4). The within-allele **tie** asserted elsewhere in this thesis rests on the two-year auto_bench corpus, where n is large enough for $p = 0.648$ to mean something (§9.1, Chapter 6 §6.10) — not on this snapshot.

The verdict here is also fragile in two ways that a single p-value hides. Of the eleven IEDB alleles, HLA-A*33:03 is single-class so its AUROC is undefined, and of the remaining ten HLA-B*27:05 has a per-allele AUROC difference of *exactly* zero; the signed-rank test's default rule discards that tied pair, so the statistic is computed on $n = 9$. On the exact $n = 9$ lattice the attainable two-sided p-values bracketing the threshold are 0.0117 and 0.0195, and the reported value is the upper one. The entire positive rank sum is carried by one allele (HLA-B*08:01, rank 3 of 9), so a single rank exchange moves the statistic one step and the result survives correction.

Second and more sharply, the outcome depends on a *convention*. Handling that tied pair by the Pratt or zero-split rules instead of discarding it gives $p = 0.015625$, which is below 0.0167 and therefore survives.[2] "Does not survive correction" thus describes a threshold sitting inside a gap of a discrete support under one particular tie-handling rule — not a margin that could be called small or large, and not a property of the data alone.

Two further caveats remain, and the first is the more serious.

The family mixes units of analysis, and its membership rule is not stated anywhere. The six tests are not commensurable. *Four* are DeLong tests treating *evaluation rows* as the unit

[2] `scipy.stats.wilcoxon` with `zero_method='pratt'` or `'zsplit'` on the ten per-allele deltas of `per_allele_iedb.csv`; the default `'wilcox'` gives $W = 3$, $p = 0.01953125$. Reproduced by `thesis_final/scripts/holm_stage1_family.py`.

— TransPHLA BA, TransPHLA EL, IEDB pooled against NetMHCpan-4.1, and IEDB against MHCflurry-2.0 ($p = 0.834$). The fifth is the IEDB per-allele Wilcoxon, whose unit is an *allele* ($n = 9$ after the tied pair is discarded). The sixth is the auto_bench per-reference Wilcoxon, whose unit is a *reference set* (46 of them over two years). Holm is valid for any finite collection of p-values, so the arithmetic is sound; what is not established is that these six are the family a reader would choose.

The construction is also not a clean reading of any one table. Chapter 6 §6.10 presents five rows carrying nine p-values; this family takes six by splitting the IEDB-versus-NetMHCpan row into its pooled and per-allele tests while collapsing each of the other rows to a single p-value. That rule is articulated here for the first time and was not pre-registered. A reader who instead took all nine p-values, or who kept one test per row, would correct over a different m. Three such families were checked — $m = 5$, 6 and 9 — and the verdict is the same in all three: the per-allele $p = 0.020$ fails its threshold under each, and no other reported conclusion moves. That robustness is reassuring but it is not the same as having specified the family in advance, and the verdict above is the one that follows from *this* family.

It is a sensitivity check, not a re-specification. The figures are computed from reported p-values rather than from a re-analysis, and they do not amend the contract of Chapter 5 §5.6, which operated at nominal significance. The Stage 2B family *is* Holm-corrected in Appendix B.

The confidence intervals do not propagate training variability. The reported intervals are not a bootstrap over "the five per-fold metrics", and should not be read as one. §5.3 resamples *row indices* with replacement and recomputes the metric once per resample, on the single ensemble score column. The real limitation is therefore different, and narrower in one respect and wider in another.

The reported intervals quantify **sampling variability in the evaluation rows**, conditional on this particular five-fold ensemble. They do not quantify variability in the *fitting* — a different seed, a different fold assignment, or a re-run of the same recipe would produce a different ensemble, and that source of variation never enters the resampling at all. The IEDB BA AUPR interval of $[0.697, 0.775]$, half-width ≈ 0.039, should therefore be read as "this ensemble, resampled rows" and not as "this recipe". The honest consequence is that the intervals are **narrower than a recipe-level interval would be**, by an amount this thesis does not measure.

Two things follow, and the second corrects a claim made elsewhere in this thesis. First, the correction we would apply is the cheap one — re-run the five-fold training under a handful of seeds and report the spread of ensemble metrics alongside the row bootstrap. It is a training cost rather than an analysis cost, and it is *not* among the three analyses §10.18 lists, which are scoped to work feasible with the released checkpoints; it should be read as a further item. Second, the ± 0.002 end-to-end AUROC variation quoted for TransPHLA in §9.10 must not be read as sitting inside the bootstrap confidence interval, which is the natural but wrong inference from two small numbers of similar magnitude. The TransPHLA EL AUROC interval has a half-width of ≈ 0.0007, so ± 0.002 is roughly three times *wider* than it. The two quantities measure different things — one is fitting variation, the other row sampling — and the fitting variation is the larger of the two on that benchmark.

The TransPHLA cutoff-date asymmetry is symmetric by construction. TransPHLA was released after the NetMHCpan-4.1 freeze, so neither predictor could have trained on it: NetMHCpan-4.1 was frozen before it existed, and DeepNeo-CL trains only on the NetMHCpan-4.1 corpus (Chapter 3 §3.4), with pair overlap re-verified. Both are evaluated frozen on identical rows under Chapter 5 §5.7. The exposure here is therefore not a leak but a difference in *era*: the two predictors were developed against different states of the field, and no statistical procedure can correct for that. We record it because it bounds how much a benchmark margin of this size can be attributed to architecture alone.

9.13 Summary

The limitations above fall into four kinds, and distinguishing them matters because they carry different weight.

Limits on what was shown. The "exceeds NetMHCpan-4.1" claim holds on TransPHLA and on deployment-safe scores only. On the static IEDB snapshot the model is behind on *both* measures — pooled cross-allele ranking (ΔAUROC -0.043, DeLong $p = 3.2 \times 10^{-8}$) and, by roughly half that margin, within allele (-0.023, Wilcoxon $p = 0.020$). The within-allele **tie** is a result of the two-year auto_bench corpus (Wilcoxon $p = 0.648$), **not** of the snapshot, exactly as Chapter 6 §6.1 states it (§9.1). Single-fold numbers are never the headline (§9.3), and the structural rescorer is withdrawn entirely because its advantage was transductive (§9.4). These are not caveats attached to a stronger claim — they define the claim.

Limits of scope. Class II MHC, post-translationally modified peptides, and TCR specificity are outside the work (§9.7, §9.8), the Stage 2B result is restricted to one held-out immunogenicity set and carries no clinical claim (§9.6), and MHCflurry-2.0's presentation channel is not benchmarked for want of a comparable interface (§9.5).

Limits of the apparatus. ESM2-35M is a probable capacity ceiling (§9.2); a larger trunk was scoped and deferred on cost-versus-risk grounds. Reproducibility is exact for *inference* from the released checkpoints but not bit-level for *retraining*, and the software stack is not fully pinned to transitive dependencies (§9.10, §9.11). Two figures remain schematic rather than generated from retained experimental output and are flagged as such wherever they appear (§9.9).

Limits we may not have seen. §9.12 lists three statistical exposures: no multiple-comparison correction across the Stage 1 comparison family (which is a mix of DeLong and Wilcoxon tests, not DeLong alone), an assumption of conditional independence between folds that is approximately but not provably true, and the cutoff-date asymmetry of the TransPHLA benchmark. The first is not merely named: §9.12 computes the Holm step-down correction and reports that the dominant TransPHLA tests survive it while the secondary comparisons are the ones a reader should weigh accordingly. The second and third are named without a magnitude, because neither can be quantified from what was retained; naming them does not discharge them. Chapter 10 sets out which of these limitations are addressable and how.

Chapter 10

Conclusion and future work

This thesis asked whether the structural information already present in the MHC class I pseudo-sequence can be used more effectively than existing pan-allele predictors use it, and whether doing so improves prediction of peptide binding affinity and antigen presentation — particularly for the long tail of rare HLA alleles, where clinical neoantigen prediction is weakest (§1.3). To answer it we built **DeepNeo-CL**: a largely frozen ESM2-35M protein language-model trunk extended with a position-resolved physicochemical encoder, peptide-conditioned cross-attention over the 34-residue pseudo-sequence, joint binding-affinity and eluted-ligand heads, long-tail imbalance correction, and a supervised contrastive refinement grouped by structurally meaningful pseudo-sequence clusters (Chapters 3 and 4). The model was trained on the official NetMHCpan-4.1 five-fold corpus (207,029 BA pairs and approximately 33 M EL pairs) and evaluated as a five-fold ensemble on three independent benchmarks — decontaminated IEDB BA ($n = 1{,}365$), TransPHLA ($n = 161{,}652$), and two years of the official IEDB weekly `auto_bench` series, which is the corpus the within-allele tie rests on — against NetMHCpan-4.1, MHCflurry-2.0 and, in Appendix D, four further predictors on identical post-cutoff rows (Chapter 3). Each comparison is reported against a declared three-condition statistical contract, together with which of its conditions were actually evaluated for that comparison — not all were, for any of them (Chapter 5 §5.6, §5.6.1, §5.6.2).

The principal finding is two-sided and is stated as such throughout. On TransPHLA, DeepNeo-CL exceeds NetMHCpan-4.1 on three of four metrics and ties on the fourth. On IEDB BA it is *not distinguished* from NetMHCpan-4.1 on within-allele ranking across two years of official weekly data — a failure to detect a difference rather than a demonstration of equivalence — and it *trails* on the small static snapshot, on pooled cross-allele ranking and within allele alike — a deficit that percentile-rank recalibration does not repair, and which is therefore reported as a representation-level limitation rather than explained away (Chapter 6 §6.4). The margins are numerically modest, and the ablation shows them to be the sum of several small, independently motivated components rather than the product of one decisive mechanism (Chapter 7).

10.1 Answers to the research questions

Chapter 1 §1.4 posed four questions. Each is answered here against the evidence, including where the answer is unfavourable or partial.

RQ1 — Structural priors. *Qualified yes.* Consuming the pseudo-sequence as a position-addressable cross-attention target is viable and is retained in the deployed model. But the

honest effect size is small: the pseudo-sequence family (concatenation plus cross-attention) contributes +0.007 TransPHLA EL AUROC after the V37 rung, and the cross-attention step in isolation only +0.002 (Chapter 7 §7.1). The mechanism earns its place by making the pseudo-sequence position-addressable for a fixed ≈ 10 per cent of the forward-pass cost (Chapter 4 §4.6.1), not by the size of its increment. That is a modest price rather than a negligible one, and the honest form of the RQ1 answer is a trade stated in both directions: a small accuracy increment bought at a real but bounded cost. A claim that cross-attention is the single largest contributor would not survive the ladder and is not made.

RQ2 — Long-tail allele transfer. *Yes, on the subgroup it targets.* Contrastive refinement grouped by pseudo-sequence cluster delivers +0.022 AUPR on the IEDB rare-allele subset ($n_{\text{train}} < 100$) without regressing common alleles, whereas grouping by allele identity (SupCon v3) regressed the tail (Chapter 7 §7.4). The structural grouping, not contrastive learning as such, is what carries the result. Note the cost: the same refinement is within noise on pooled TransPHLA EL (−0.001 AUROC), so this contribution is a subgroup contribution and is reported as one.

RQ3 — Length extremes. *Partly, and the limitation is real.* A single pan-length model with position-resolved physicochemical features — and *without* any length-aware imbalance correction, since neither length-conditioned variant tested was promoted (§7.2) — scores competitively across 8–11mers without separate per-length models, leading on 8mers and 11mers and trailing on the dominant 9mer slice (Chapter 6 §6.3). The gains at the non-canonical lengths offset the 9mer deficit rather than overturning it — the frequency-weighted per-length AUPR delta is slightly negative — so the honest answer is that one model *suffices* over this range, not that it wins over it. And the thesis cannot answer the question at all for 12–14mers: the benchmark is 8–11mers only, so no claim is made about the longer tail that actually motivated the gap, and Chapter 9 records this as an open limitation rather than a settled result.

RQ4 — Measurement. *Yes on TransPHLA AUROC, tied on IEDB within-allele, no on pooled IEDB.* The two TransPHLA *AUROC* gains satisfy both conditions that were evaluated; condition (a)'s paired interval was never computed for any comparison, so no verdict in this thesis satisfies all three, and the AUPR gains — including the BA AUPR margin often quoted beside them — lie outside the contract's reach entirely, so for those no condition was evaluated (§5.6.2). The IEDB within-allele comparison is a statistical tie over two years of auto_bench ($p = 0.65$), and the pooled static-snapshot comparison fails in DeepNeo-CL's disfavour. The contract should not be credited with the withdrawal of the crystal-template pocket rescorer: it had scored that component favourably, and what disqualified it was a later judgement about fitting on evaluation labels (§5.6.1, Chapter 7 §7.5, §7.10). The withdrawal cost the thesis its most favourable headline, which is a fact about the thesis's honesty and not about the contract's.

Taken together the four answers support a narrow conclusion rather than a broad one, and §10.3 states it in the form the evidence will bear.

10.2 Summary of contributions

The research question each contribution answers is given in brackets; §10.1 states how far the evidence carries each one.

- **C0 — The DeepNeo-CL framework.** The principal outcome of this thesis is the predictor itself: an integrated, trained and independently evaluated peptide–MHC-I binding and presentation model, released as a five-fold ensemble with its weights, prediction arrays, statistical outputs and a CPU forward-pass validator (Appendix A). The contributions C1–C4 below are the components that distinguish it; C0 is the artefact that combines them and is the thing a reader can run. *[RQ1–RQ4]*

- **C1 — Architecture.** A peptide-conditioned cross-attention layer over the 34-residue MHC pseudo-sequence inside an ESM2-35M transformer backbone. In the cumulative ablation the cross-attention step contributes +0.002 TransPHLA EL AUROC in isolation, and the pseudo-sequence *family* — concatenation plus cross-attention — contributes +0.007 measured after the V37 rung. The +0.012 sometimes quoted here is the cumulative span from the peptide-only ESM2 baseline and includes V37's own +0.005, so it is not the family's contribution and is not claimed as one (§7.1). The mechanism costs a fixed ≈ 10 per cent of the forward pass (§4.6.1). *[RQ1]*

- **C2 — Training method.** SupCon v2 — a supervised contrastive auxiliary loss on the *pseudo-cluster* label (12 clusters), fp32-wrapped and NaN-guarded — regularises the embedding space along the structurally meaningful axis and delivers +0.022 AUPR on the IEDB rare-allele subset. It is within noise on pooled TransPHLA EL, so this is a subgroup contribution. *[RQ2]*

- **C3 — Deployment-safe BA+EL fusion.** A label-free rank-average of the two frozen heads (BA + EL) lifts pooled TransPHLA BA from 0.9546 to 0.9694, above NetMHCpan-4.1's published BA output of 0.9638, with no eval-set fitting (paired DeLong $z = 22.9$). **That comparison is fused against unfused**, and the symmetric check matters: applying the same rank-average to NetMHCpan-4.1's own two heads gives 0.9697 against DeepNeo's 0.9694 (§6.5.2) — a gap of 0.0003 in NetMHCpan's favour, for which no paired test is reported, so the like-for-like comparison is **not resolved either way** rather than demonstrably tied. The honest claim is that the recipe is a real elevation of this model's BA channel and not a demonstration of architectural superiority — fusion helps DeepNeo-CL more because its BA head sits 0.0151 below its EL head where NetMHCpan's gap is 0.0044. An exploratory crystal-template pocket rescorer (Chapter 7 §7.5) showed transductive in-sample gains only and is excluded from all deployment claims. *[RQ4]*

- **C4 — Methodology.** A declared three-condition statistical contract (bootstrap CI + DeLong + per-allele Wilcoxon, with its provenance stated at §5.6.1 and the gaps in its own application at §5.6.2), exact pair-level decontamination of the external benchmark, and a reproducibility vault released alongside the thesis. The decontamination is *exact-pair* and not near-duplicate-aware, and the sensitivity analysis that would bound the difference was specified and not run (§3.9, §9.1). *[RQ4]*

10.3 The single defensible claim, as discharged

Chapter 1 §1.5 stated one claim and undertook to defend it. This section reports what the evidence did to it, rather than restating it. Reproducing that block verbatim would let a conclusion assert exactly what an introduction had promised, with no trace of the intervening 150 pages.

What survived unchanged. The TransPHLA half. DeepNeo-CL exceeds NetMHCpan-4.1 on BA AUROC and AUPR after label-free BA+EL fusion (0.9694/0.9705 vs 0.9638/0.9642; DeLong $z = 22.9$) and on EL AUROC (0.9697 vs 0.9682, $p = 1.8 \times 10^{-8}$), and ties on EL AUPR (0.9714 vs 0.9716). These margins are small and they are resolvable at $n = 161{,}652$; both facts belong in the claim.

What had to be weakened. Two things, and in the same direction. First, "statistically equivalent" was not available: the auto_bench within-allele result is a *failure to distinguish* (Wilcoxon $p = 0.65$), and a non-significant test is not evidence of equivalence (§9.12). Second, "trails only on pooled" was not true: on the 1,365-row static snapshot DeepNeo-CL trails on *both* measures — pooled (ΔAUROC -0.043, DeLong $p = 3.2 \times 10^{-8}$; ΔAUPR -0.0651, $\approx 3.3\,\sigma$) and within allele (-0.023, $p = 0.020$). The pooled IEDB deficit is a measured loss, not a limit of sample size.

What was withdrawn outright. The crystal-template pocket rescorer and every number derived from it: its cross-validation was fitted on the evaluation set and it collapses inductively, IEDB BA AUROC 0.862 → 0.442 (§9.4). The thesis is weaker for its removal and correct for it.

The claim as it now stands. DeepNeo-CL — an ESM2-35M pMHC-I predictor with peptide×HLA pseudo-sequence cross-attention and SupCon v2 representation refinement,[1] evaluated as a five-fold ensemble with no eval-set fitting — exceeds NetMHCpan-4.1 on the external TransPHLA benchmark by small but resolvable margins, is not distinguished from it on within-allele IEDB ranking over two years of auto_bench, and is behind it on the static IEDB snapshot on both pooled and within-allele measures. Paired tests in Chapter 6 §6.10 and the artefacts in the milestone vault[2] support each part of that sentence, including the unfavourable parts.

10.4 What the milestone vault enables

The milestone vault[3] ships:

1. The exact code (byte-for-byte) that produced every result.
2. The 5-fold ensemble weights with MD5 integrity hashes.
3. The raw `y_true / y_pred` arrays for every reported metric.
4. The statistical-summary log with bootstrap CI, DeLong, Wilcoxon.
5. The publication-grade figures in 300+ DPI PDF and PNG.
6. The single-pair forward-pass validator (under 2 minutes CPU).

The vault is the operational contract of this thesis: any reader who disagrees with a number can verify or refute it without needing the training cluster. To our knowledge it is among the most complete artefacts released alongside a pMHC-I predictor thesis.

[1] Stage 1 pipeline as defined in `docs/model/STAGE_NAMING.md` v2.
[2] `milestone_netmhcpan_beaten/`.
[3] `milestone_netmhcpan_beaten/`.

10.5 Honest reflection on what this thesis is not

The contribution defended here is small in absolute terms — one new top-tier metric on one external benchmark — but it is operationally meaningful. This thesis is **not**:

- A clinical neoantigen-discovery study. The headline numbers are on binding/presentation plus Stage 2B BigMHC immunogenicity; clinical efficacy still requires Stage 2A TCR-conditioned modelling and prospective patient validation (out of scope).
- A breakthrough architecture. No single rung on the cumulative ablation ladder contributes more than +0.005 AUROC, and the largest step is matched by two others (Chapter 7 §7.1). The win is composed of multiple small improvements along orthogonal axes.
- A universal architectural recommendation. The 7 retained components compose linearly in our pipeline; we have not ablated their ordering.
- An IEDB win of any kind. On the two-year auto_bench corpus DeepNeo-CL is *not distinguished* from NetMHCpan-4.1 on the deployment-fair per-reference metric (Wilcoxon $p = 0.65$), which is a failure to detect a difference and not a tie demonstrated. On the 1,365-row static snapshot it is **behind on both** measures — pooled (-0.043) and within allele (-0.023) — so "trails only on pooled" is not available either (§10.3).

10.6 What this thesis claims for the field

We make two structural arguments to the pMHC-I prediction community:

1. **Methodological rigour is overdue.** As discussed in Chapter 2 §2.4, a qualitative review of recent transformer-based pMHC-I predictors suggests that confidence intervals on the headline metric, k-fold ensemble reporting, and strict benchmark decontamination are inconsistently practised. This thesis argues that the same numerical result could not be reported in the recent literature's prevailing rigour standards, and that the field would benefit from raising those standards. We frame this as a qualitative argument; a pre-registered systematic review is beyond the scope of the thesis.
2. **The reproducibility vault should be a default deliverable.** For every new pan-allele predictor: ship the weights, the raw predictions, the bootstrap CIs, and a single-pair validator. The vault format proposed here is one template; the principle is more important than the specific implementation.

10.7 The Stage 1 -> Stage 2 bridge

Stage 1 closes the **BA** and **EL** halves of the biological antigen-recognition pathway: *does the peptide bind?* and *is it presented on the cell surface?* The next biological step – *is the presented pMHC complex recognised by a CD8+ T-cell receptor?* – is the **immunogenicity (IM)** question, addressed by Stage 2B in Appendix B. In this thesis, Stage 2B is **not** a separate predictor on top of a generic feature extractor: it consumes the same 256-dimensional fused hidden representation that Stage 1 supervises jointly with BA and EL labels, so the immunogenicity head reads the

binding-pocket geometry, presentation likelihood, and per-allele anchor compatibility that Stage 1 already encoded. In this sense Stage 1 and Stage 2B form **one end-to-end DeepNeo-CL system** that returns BA + EL + IM scores from a single forward pass over (p,h).

At the same time, this bridge has a clear boundary: pMHC-only representation can travel into immunogenicity transfer (Stage 2B), but full clinical specificity ultimately requires explicit TCR-conditioned modelling (Stage 2A).

The current work does not extend to class-II antigen presentation, nor to explicit TCR-sequence modelling (Stage 2A, future work).

10.8 Where the two stages leave the pipeline

Stage 1 establishes binding and presentation scores competitive with NetMHCpan-4.1 under the declared evaluation contract; Stage 2B adds an immunogenicity head on the same trunk without retraining the backbone. The resulting pipeline returns BA, EL, and IM scores in one forward pass, with each stage evaluated against external comparators on its own task. The milestone vault[4] provides the weights, predictions, and validators needed to reproduce every reported number without access to the original training cluster.

Chapter 9 stated what this thesis does not establish. This chapter sets out which of those gaps can be closed, by what means, and at what cost — and distinguishes them from the ones that cannot. The organising principle is feasibility rather than ambition: each proposal below states the work required, the resources it needs, and the specific limitation or research question it would resolve, so that a reader can judge whether it is a next step or a research programme.

What remains to be done

The outstanding work falls into three bands.

Directly addressable extensions of this thesis. Work that reuses the trained model and the existing evaluation apparatus, and could be completed without new data collection: a leave-one-out ablation to complement the cumulative ladder (§10.18.1), per-pocket gradient attribution over the trained model to test the anchor-chemistry mechanism proposed in Chapter 8 §8.1 (§10.18.2), an unfrozen-trunk transfer-learning comparison against the frozen-trunk deployment configuration (§10.18.3), and multiple-comparison correction applied uniformly across all reported tests rather than only the Stage 2B family (Chapter 9 §9.12). These are the items a reviewer is most likely to ask for, and none requires new experimental data.

Scoped follow-on projects. Work that needs new data, new training runs, or substantially more compute, and that would constitute a project in its own right: TCR-conditioned Stage 2A modelling (§10.9, §10.12), the ESM2-150M redux that would test the suspected capacity ceiling (§10.10), class II extension (§10.11), distillation for clinical throughput (§10.13), an AlphaFold3-based replacement for the withdrawn structural rescorer (§10.14), adversarial robustness (§10.15), and the engineering and tooling work of §10.17.

Out of reach of this line of work. Clinical validation, and any claim about post-translationally modified peptides, require experimental immunology rather than better modelling and are

[4] `milestone_netmhcpan_beaten/`.

not framed here as future computational work.

§§10.9–10.17 develop each proposal, and §10.18 ranks them by return on effort.[5]

10.9 Stage 2A TCR-conditioned modelling (highest-ROI)

The single highest-ROI next experiment is **Stage 2A TCR-conditioned modelling**. Stage 2B immunogenicity transfer is now reported in Appendix B on the BigMHC `im_test` benchmark; the largest remaining gap to clinical applicability is explicit TCR-specificity modelling on peptide:MHC:TCR triplets.

10.9.1 Proposed protocol

1. **Data**: de-duplicated VDJdb + McPAS-TCR + IEDB T-cell assay union, pair-decontaminated against Stage 1/2B training pairs.
2. **Method**: load the Stage 2B-refined trunk, freeze it, enable the TCR branch (`tcr_dim > 0`), and train a lightweight triplet head on peptide:MHC:TCR tuples.
3. **Evaluation**: strict held-out split with paired comparison against NetTCR-2.0, TITAN, ERGO-II, and pMTnet on the same rows.
4. **Cost estimate**: $\approx$ 4–6 GPU-hours.
5. **Statistical contract**: same three-condition contract as the binding work.
6. **Risk**: TCR datasets are noisy and sparse across HLA alleles; reported gains can collapse under strict pair-level decontamination.

If this experiment succeeds under strict paired evaluation, it becomes the headline extension of this thesis programme. Even if it fails, the result is still informative: pMHC-only features (Stage 2B) are likely near their ceiling without explicit TCR conditioning.

10.10 ESM2-150M Stage 1 redux (deferred, low risk)

10.10.1 Rationale

ESM2-35M is a capacity ceiling. ESM2-150M (4× parameters) might deliver a small IEDB lift at ~30 GPU-hours cost. The Phase 1 ablation (Chapter 7 §7.6) measured a delta within 95% CI on the 35M $\rightarrow$ 35M-with-MLM-CPT comparison, which is not the same as a 35M $\rightarrow$ 150M comparison.

10.10.2 Risk-cost calculus

- Expected gain: +0.001 to +0.005 IEDB AUPR (within current CI).
- Compute cost: 30 GPU-hours for full 5-fold rebuild.
- Risk: backbone-size scaling on long-tail data may not be linear.

[5] The underlying strategy notes are `docs/model/IMPROVEMENT_STRATEGY.md` and §7 of the experiment history, in the source repository.

We defer this because (a) the TransPHLA win does not need the larger backbone, and (b) the subsequent manuscript can ship without it. If 30 GPU-hours becomes available later in this project, this is the next compute to spend.

Note: this experiment was historically called "Phase 1 redux" in internal documentation; under the canonical taxonomy of the project's stage naming document[6] it is a Stage 1 re-training with a larger backbone, retaining all other recipe elements.

10.11 Class II MHC extension

The DeepNeo-CL methodology — peptide × pseudo cross-attention and SupCon contrastive regularisation — should transfer to class II MHC (HLA-DR, -DQ, -DP) with adjustments for:

- longer peptide length (typically 13–25 residues);
- open binding groove (no length cap);
- different anchor geometry (P1, P4, P6, P9 are the canonical class II anchors);
- different reference predictor (NetMHCIIpan).

The same statistical-rigour standards (bootstrap CI, 5-fold ensemble, strict decontamination) should apply. This is a natural follow-on thesis or a stand-alone paper.

10.12 Peptide:MHC:TCR triplet modelling (Stage 2A)

Current TCR status. DeepNeo-CL operates on the *binary* pMHC interaction. The source code contains a small amount of optional architectural scaffolding for a future TCR branch (the model's summary-attention module[7] accepts an optional `tcr_emb` argument), but the scaffolding is **dormant** at training and inference time in every reported benchmark in this thesis. No TCR training data has been ingested; no TCR-specific training has been performed; no benchmark comparison against industry TCR predictors (NetTCR-2.0, TITAN, ERGO-II, pMTnet, tcrdist3) has been executed. The detailed Stage 2A status is documented in the project repository's TCR status memo, indexed in the Project Documentation Index (Appendix C).

Why deferred. Clinical neoantigen efficacy ultimately depends on the **triplet** problem: which TCRs in a patient's repertoire will recognise a given pMHC. This is a central open problem in the field and is currently bottlenecked by the availability of paired pMHC–TCR training data (VDJdb $\approx 10^4$ confident pairs vs. 207 K NetMHCpan-4.1 BA pairs and 33 M EL pairs). Robust transfer from the binding task to the triplet task is not guaranteed at this scale.

Future Stage 2A protocol. If the TCR-binding-prediction community produces a sufficiently large paired benchmark, the natural next step is:

1. **Data**: VDJdb + McPAS-TCR + IEDB T-cell assay (de-duplicated, pair-decontaminated against NetMHCpan-4.1 training).
2. **Backbone**: freeze the Stage 1 SupCon refinement trunk; activate the dormant TCR scaffolding by setting `tcr_dim > 0`.

[6] `STAGE_NAMING`.

[7] `src/models/deepneo_cl.py`, class `SummaryAttention`.

3. **Architectural device**: a cross-attention layer with the TCR CDR3α/β as the third sequence.
4. **Benchmark**: head-to-head against NetTCR-2.0, TITAN, ERGO-II, pMTnet on a shared held-out IEDB T-cell assay slice under the three-condition statistical contract.
5. **Realistic ceiling**: TCR-specificity AUROC is probably 0.6–0.75 (vs. the 0.97 ceiling on pMHC binding/presentation). A "win" would be a modest absolute number by binding-prediction standards.

This is scoped as a follow-on revision or stand-alone paper, not as part of the present thesis.

10.13 Distillation for clinical-pipeline throughput

DeepNeo-CL at inference throughput is $\approx$ 200 pairs/sec on a single A100. A whole-exome neoantigen pipeline that needs $\approx 10^5$ predictions per patient finishes in $\approx$ 8 minutes — perfectly fast for offline analysis. For real-time clinical decision support, distillation into a smaller student network (e.g. a 6-layer transformer) could deliver 10× speedup. This is engineering work, not research.

10.14 AlphaFold3-API features for a next structural-reranker experiment

The current pocket-reranker experiment uses crystal-template descriptors. An obvious extension is a next structural-reranker experiment — replace the crystal-template features with AlphaFold3-predicted structure features (full HLA heavy chain, not just pseudo-sequence). Risks:

- AlphaFold3 API quotas and licensing.
- Whether the AF3 pHLA confidence is high enough (the Phase 3 v1 (rejected ESMFold reranker) ESMFold ablation said no for pseudo-only; AF3 with full heavy chain may be different).
- Per-pair latency.

If feasible, this next structural-reranker experiment would close the structural-prediction loop and deliver an end-to-end neural pipeline (no crystal-template prior).

10.15 Adversarial robustness

We have not stress-tested DeepNeo-CL on:

- peptides outside the canonical 20-amino-acid alphabet (e.g. selenocysteine, pyrrolysine, modified residues);
- peptides at the boundary of the length envelope (7-mers, 15-mers);
- allele rare-variants outside the 340-vocabulary training set.

A formal robustness study (small perturbation, allele substitution, length extrapolation) is a publishable methodology paper.

10.16 Vault-as-template: methodological adoption

The milestone vault[8] is, to our knowledge at submission time, among the most complete peer-review-grade reproducibility artefacts released alongside a pMHC-I predictor thesis. We hope it becomes a community template. Concrete recommendations for the field:

1. Every new pan-allele predictor should ship a vault with:

- frozen code + MD5 checksums;
- 5-fold ensemble weights;
- raw `y_true / y_pred` arrays;
- bootstrap CI + DeLong test report;
- a single-pair forward-pass validator.

1. Benchmarks should be reported on the same external rows as existing literature, not on author-curated splits.
2. Single-fold numbers should generally not be the headline.
3. IEDB BA (at $n = 1{,}365$) should be reported *alongside* a larger external benchmark such as TransPHLA rather than in isolation.

The Open Science framework analogue would be: pre-register the benchmark, ship the vault on Zenodo with a DOI, and report only results that clear the three-condition contract.

10.17 Engineering and tooling

Several small engineering items would improve thesis-revision and post-publication operations:

- Persist per-fold training-loss / validation-AUPR logs at fold-level granularity (currently lost between Stage 1 and the vault, see Chapter 9 §9.9).
- Persist the per-pair shared representations the SupCon loss acts on, to enable a literal t-SNE for Figure 7.2 (currently schematic).
- Add a Bonferroni-corrected version of the DeLong p-values table for full methodological belt-and-braces.
- Mint a Zenodo DOI for the vault directory (currently lives on GitHub master only).

[8] `milestone_netmhcpan_beaten/`.

10.18 Analyses that would strengthen the present claims

Four analyses are distinguished from the proposals above because they would not extend the model's scope — they would test or sharpen claims the thesis already makes. They differ in cost, and each section states its own reason for not being run; the common constraint is stated here.

One reason suggests itself: that the deployed model and its reported numbers are frozen, so a re-run altering a graded result would need reporting as a new result. **That rationale does not fit all four**, and offering it as the reason would make a scoping decision look like a methodological necessity. Two of them would alter nothing reported — the attribution map (§10.18.2) is forward and backward passes over frozen weights, and the leave-one-out grid (§10.18.1) trains additional variants and leaves the deployed numbers untouched. One produces a separate comparator by training unfrozen folds (§10.18.3). And one, the homology sensitivity analysis (§10.18.4), *would* change a reported comparison — which is precisely why it should be run, not a reason to avoid it.

The constraint that does apply to all four is scope: the integrated thesis layer is downstream-only and introduces no result that does not already exist upstream, so new computation belongs in a fresh Stage 2 run with its own provenance rather than as a late addition here. That is a boundary, not a defence. Cost is a second and independent constraint, and it binds differently in each case — §10.18.1 is a 105 GPU-hour grid while §10.18.4 needs no model inference at all — so "feasible with the released checkpoints" is true of the analyses' *inputs* and not of their budgets.

10.18.1 Leave-one-out ablation to complement the cumulative ladder

The ablation of Chapter 7 is cumulative: each component's reported contribution is its marginal effect *given* the components already present (§7.1). A cumulative design answers "what did adding this buy?" but not "what would removing this cost?", and the two can differ when components are partially redundant — a component that looks small when added late may be substantial when removed from the complete model, because another component has absorbed its function. The complementary experiment is a leave-one-out grid: retrain the full configuration k times, each time with one component disabled, and report the degradation. At seven retained components and roughly three GPU-hours per base fold the full grid is approximately 105 GPU-hours for five-fold ensembles — but that figure overstates what the useful experiment costs. The single most valuable cell is one component: whether cross-attention's small *marginal* contribution (+0.002 AUROC) understates its *necessary* contribution — the attribution in this thesis most open to challenge — which is about fifteen GPU-hours at five folds. And one cell is already in hand: the configuration without SupCon v2 is the penultimate rung of §7.1, so that leave-one-out costs nothing further. The experiment is two-sided, and the thesis would report it either way: the result that would most strengthen the architectural claim is that removing cross-attention costs more than its +0.002 marginal suggests, and the result that would most weaken it is that removing it costs little, i.e. that another component has absorbed its function and it is closer to redundant than the cumulative figure implies.

10.18.2 Per-pocket gradient attribution over the trained model

Chapter 8 §8.1 proposes that the model's advantage is mediated by anchor chemistry at the P2 and PΩ positions, and supports this with the per-length breakdown and the V37 ablation — both indirect evidence. The direct test is an attribution map: compute gradients of the output with respect to each

pseudo-sequence position, aggregated by binding-groove pocket, and ask whether the positions the model actually relies on are the pockets that structural biology identifies as anchor-contacting. This is inexpensive, requiring only forward and backward passes over the released checkpoints with no retraining.

One reason has to be ruled out explicitly, because it is the tempting one: that a negative result would weaken a Chapter 8 interpretation. That is not a reason to leave an experiment unrun — it is the outcome-selection bias this thesis claims to avoid, and stating it as a justification is worse than the omission it was meant to excuse. A cheap test whose result might be unwelcome is precisely the test that should be run.

The actual reason is one of **scope**, and it binds the whole thesis rather than this analysis: every number reported here is derived from pre-existing frozen artefacts, and a new computation — even a cheap one — would introduce a result that exists nowhere upstream of this document. The place for it is a fresh Stage 2 run with its own provenance, not a late addition to a frozen chapter.

The cost of that decision is stated rather than absorbed. The anchor-chemistry mechanism of §8.1 rests on two pieces of *indirect* evidence — the per-length breakdown and the V37 ablation — and this would have been the direct test. Until it is run, §8.1 should be read as a proposal with no direct support, and a reader who discounts it entirely is not misreading the thesis. No quantitative per-pocket attribution is claimed anywhere in this work.

10.18.3 Unfrozen-trunk transfer learning as a comparator

The deployed configuration freezes the ESM2-35M trunk except for its top two layers, a choice made because a full fine-tune produced no benchmark gain for roughly thirty GPU-hours (Chapter 4 §4.3). That finding is recorded as a single Phase 1 experiment and deserves a fuller treatment: the open question is not whether full fine-tuning helps on average but whether it would help specifically on the pooled cross-allele IEDB axis where the model trails (Chapter 6 §6.4), since that deficit is argued to be representation-level rather than calibration-level. The experiment is a matched comparison between frozen-trunk and unfrozen-trunk training on identical folds, evaluated under the same contract, with the cost in both compute and in the loss of the frozen trunk's parameter-efficiency argument (§4.6.4) reported alongside any gain. Parameter-efficient adaptation such as low-rank adapters is the cheaper middle case and is the variant we would run first.

10.18.4 The homology-reduction sensitivity analysis, which was specified and not run

This one is listed differently from the other three, because it is not an analysis that *would* strengthen a claim — it is an analysis this thesis described as performed and reported a result for, and which does not exist. Chapter 3 §3.9 and Chapter 6 §6.11 record the retraction; this section records what running it would take.

The design is in §3.9: for each surviving IEDB pair, flag it when any training peptide of the same length against the same allele shares a contiguous match of at least $L-1$ residues, then recompute the headline comparison on the unflagged subset. It requires no model inference and no retraining — only the cleaned benchmark, the training pair set, and a string-matching pass — so it is the cheapest item in this chapter, cheaper than §10.18.2. Two thresholds are worth reporting, $L-1$ and $L-2$, because the first bounds near-identity and the second bounds the single-substitution

neighbourhood that matters most for a 9-mer.

Two further things should be reported with it, and were not available for this thesis. First, the number of pairs each threshold removes, which the committed metadata does not record (§3.5.2). Second, the effect on *both* predictors, since both train on the NetMHCpan-4.1 corpus and near-duplicate retention inflates whichever has seen the neighbour — a flagged-subset comparison that reported only DeepNeo-CL's movement would be the same one-sided presentation this chapter criticises elsewhere.

10.19 Summary of priorities

In decreasing order of expected return:

1. **Stage 2A TCR-conditioned modelling** (highest-ROI next experiment after Stage 2B completion).
2. **Vault-as-template community advocacy** (publishable methodology paper).
3. **Class II MHC extension** (full follow-on thesis or paper).
4. **Distillation for clinical throughput** (engineering).
5. **AlphaFold3-driven structural reranker follow-up** (technical).
6. **ESM2-150M Phase 1 redux** (cost-prohibitive at current project budget).
7. **Stage 2B external-dataset expansion** (e.g., IEDB T-cell union under the same paired statistical contract).
8. **Adversarial robustness study** (orthogonal methodology paper).

Appendix A

Reproducibility appendix

This appendix is the single-document recipe for an independent reader to reproduce the *base-ensemble* numerical claims of the thesis from a read-only artefact (the milestone vault) plus, optionally, a single A100-class GPU for a full from-scratch retrain. It is **not** the whole thesis: the fused headline row, the four-predictor comparison and the two-year `auto_bench` figures come from artefacts outside the vault, which Chapter 5 §5.10 names.

A.1 Layered reproducibility model

Four layers of reproducibility, in increasing cost:

Layer	What it reproduces	Cost	Path
L1	Read the base-ensemble numbers and the figures	0	Read `thesis_assets/thesis_stats.txt` and the 12 figure PDFs
L2	Run a single forward pass to confirm checkpoint integrity	< 2 min CPU	`python scripts/validate_vault.py`
L3	Re-derive `thesis_stats.txt` from the 10 checkpoints + benchmark CSVs	~ 15 min on A100	`python scripts/extract_zero_rerun_assets.py`
L4	Re-derive everything from the upstream NetMHCpan-4.1 corpus	~ 17.5 GPU-hours on A100	full pipeline (§A.6)

Table A.1: The four reproducibility layers, in increasing cost. L1 and L2 need no GPU; L3 re-derives the statistics file and L4 retrains from the upstream corpus. Note that L2 confirms checkpoint integrity and verifies no reported metric — re-deriving a number is L3.

A reviewer who wants to defend or attack any claim in this thesis should pick the lowest sufficient layer.

A.2 The milestone vault (`milestone_netmhcpan_beaten/`)

```
milestone_netmhcpan_beaten/
+- README                                  vault overview
+- README                                  index, commit SHA, canonical baseline
+- code/                                     exact training/eval scripts
+- weights/                                  10 checkpoints with MD5 hashes
|   +- supcon_5fold/c00{0..4}/multitask_ba_el.pt
|   +- e13_5fold/c00{0..4}/multitask_ba_el.pt
|   +- checksums.md5
|   +- architecture_metadata.json
+- docs/
|   +- CONTEXT_RESUME                      single-entry resume
```

```
|   +-- THESIS_CHAPTER_DRAFT                Chapter
↪  \ref{ch:C04_datasets:chapter-4-datasets-and-preprocessing} source
|   +- THESIS_WRITING_PLAN                 schedule + defence worksheet
|   +- methodology_summary                 CLI playbook
|   +- references.bib                         BibTeX seed
|   +- chapters/C0{1..8}_*                 chapter drafts
+- thesis_assets/
|   +- thesis_stats.{txt,json}                literal stats
|   +- *_raw_predictions.npz                  raw y_true + y_pred
|   +- per_allele_{iedb,transphla}.csv        per-allele breakdown
|   +- per_length_{iedb,transphla}.csv        per-length breakdown
|   +- dummy_forward_pass.json                validator cache
|   +- raw_benchmark_logs/                    reranker JSONs
|   +- figures/                               12 figures × {pdf,png}
+- data_blueprint/                            4 protocol docs (see caveat below)
+- scripts/
    +- copy_into_vault.sh                     Step 1: build vault payload
    +- extract_zero_rerun_assets.py           Step 3: build stats + arrays
    +- make_figures.py                        Step 4: regenerate figures
    +- validate_vault.py                      Step 5: integrity check
    +- generate_thesis_assets.py              deterministic synthetic generator
```

Total disk footprint after vault build: **1.7 GB**.

One caveat on the protocol documents. The vault was frozen on 2026-05-20 and is not edited by later corrections, so two of its documents assert things this thesis has since withdrawn. `homology_reduction_notes` reports a result for the homology-reduction sensitivity analysis and names a flagging script; neither the result nor the script exists, and Chapter 3 §3.9 and Chapter 6 §6.11 record the retraction. The vault `README` likewise refers to an MD5 manifest that was never generated (§A.12). A reader following the vault's own pointers should take the thesis chapters as authoritative where the two disagree.

A.3 L1 — Read-only reproduction

For a reviewer who wants only to read and verify numbers:

```
git clone git@github.com:KevinSimple/deepneo_cl.git
cd deepneo_cl
cat milestone_netmhcpan_beaten/thesis_assets/thesis_stats.txt
ls milestone_netmhcpan_beaten/thesis_assets/figures/
```

The *single-head* benchmark cells of Chapters 6 and 7 trace to a row in `thesis_stats.txt`. Not every numerical claim in those chapters does, and the exceptions include the thesis's principal result — the file contains no fusion rows at all. Chapter 5 §5.10 lists what lies outside this layer. Every figure referenced in the chapters lives in `figures/` at 300 DPI in both PDF and PNG.

A.4 L2 — Integrity check (< 2 minutes CPU)

```
cd milestone_netmhcpan_beaten
python scripts/validate_vault.py
```

This script performs three independent checks:

1. **MD5 hashes** of every checkpoint match `weights/checksums.md5`.
2. **Architecture metadata** in `weights/architecture_metadata.json` is consistent with the loaded checkpoint state-dict shapes.
3. **Single-pair forward pass** on a hardcoded dummy `(SIINFEKL, HLA-A*02:01)` reproduces the cached value in `thesis_assets/dummy_forward_pass.json` to within 10^{-4} absolute tolerance.

Exit code 0 on success. Runs in < 2 minutes on CPU; no network access required.

A.5 L3 — Stats re-derivation (~ 15 minutes on A100)

To reproduce `thesis_stats.txt` from the 10 checkpoints + benchmark CSVs:

```
cd /workspace/deepneo_cl   # the persistent RunPod volume
export PYTHONUTF8=1
./venv/bin/python milestone_netmhcpan_beaten/scripts/extract_zero_rerun_assets.py
```

This loads the 5-fold ensemble checkpoints, predicts on IEDB + TransPHLA + TESLA using deployment-safe scores (base ensemble and label-free BA+EL fusion), computes bootstrap CIs + DeLong + Wilcoxon, and writes the regenerated `thesis_stats.{txt,json}` to `thesis_assets/`.

To regenerate figures from the regenerated stats:

```
./venv/bin/python milestone_netmhcpan_beaten/scripts/make_figures.py
./venv/bin/python milestone_netmhcpan_beaten/scripts/_regen_figs_from_real.py
```

A.6 L4 — Full from-scratch retrain (~ 17.5 GPU-hours)

For a reviewer who wants to retrain DeepNeo-CL end-to-end:

A.6.1 Environment

```
cd /workspace/deepneo_cl
python3 -m venv venv
./venv/bin/pip install -r requirements.txt
./venv/bin/pip install --upgrade torch torchvision torchaudio \
  --index-url https://download.pytorch.org/whl/cu124
export PYTHONUTF8=1
export HF_HOME=/workspace/hf_cache
export HF_HUB_CACHE=/workspace/hf_cache/hub
```

A.6.2 Stage 0 — Data preparation

```
./venv/bin/python scripts/stages/stage0_prepare_netmhcpan41_data.py \
  --source data_raw/netmhcpan41 \
  --out data_proc/stage1_traindata_netmhcpan41
```

Outputs `binding_cv/folds/c00{0..4}/{train,test}.csv` and `presentation_cv/folds/c00{0..4}/{train,test}.csv`. Wall-clock: 10 min CPU.

A.6.3 Stage 1 — E13 5-fold

```
for k in 3 4; do
  ./venv/bin/python scripts/stages/stage1_multitask_train_ba_el.py \
    --config configs/audit_xattn_base_c00${k}.yaml
done
```

Wall-clock: ~3 h per fold, ~15 h total on a single A100 for all five folds.

Only folds c003 and c004 can be reproduced from the repository. The configurations for c000–c002 were not committed — those folds were run from earlier-named files that no longer exist in the history — so the loop above is restricted to the two that do. Chapter 3 §3.24 states the consequence: inference from the released checkpoints is exact and unaffected, but byte-for-byte retraining of three of the five base folds is not possible from this repository alone.

A.6.4 SupCon refinement — v2 continuation

All five SupCon folds run from the single committed configuration, which carries `fold: c000` natively; folds c001–c004 override it on the command line:

```
# c000 runs from the config as committed
./venv/bin/python scripts/stages/stage1_multitask_train_ba_el.py \
  --config configs/phase2_supcon_v2_c000.yaml

for k in 1 2 3 4; do
  ./venv/bin/python scripts/stages/stage1_multitask_train_ba_el.py \
    --config configs/phase2_supcon_v2_c000.yaml --fold c00${k} \
    --resume <base checkpoint for fold c00${k}>
done
```

Wall-clock: ~25 min per fold, ~2 h total. The continuation resumes from the corresponding Stage 1 base checkpoint, so this step is reproducible only for the folds whose base run is reproducible — see §A.6.3. The exact resume paths are in `scripts/runpod/tasks/run_phase2_supcon_5fold.sh`, and two of them carry the architecture-mismatch caveat recorded in the header of `configs/audit_xattn_base_c003.yaml`.

A.6.5 Structural pocket rescorer (withdrawn experiment)

These commands reproduce the crystal-template rescoring experiment that is **withdrawn as transductive** (Chapter 7 §7.5, Chapter 9 §9.4). They are provided for provenance only; the experiment contributes no deployment result and no headline number.

```
./venv/bin/python scripts/structure/precompute_pocket_features.py \
  --input data_proc/iedb_benchmark/iedb_ba_clean.csv \
  --out data_proc/structure_cache/iedb_pocket_features.npz

./venv/bin/python scripts/structure/precompute_pocket_features.py \
  --input data_proc/transphla_benchmark/transphla_independent.csv \
  --out data_proc/structure_cache/transphla_pocket_features.npz

# Train and evaluate the reranker on each channel
```

```
for baseline in supcon_5fold e13_5fold; do
  for task in ba el; do
    ./venv/bin/python scripts/structure/train_eval_struct_reranker.py \
      --benchmark transphla --baseline ${baseline} --task ${task}
  done
done
```

Wall-clock: ~3 min for caches + 5 min × 4 channels for reranker.

A.6.6 Evaluation

```
./venv/bin/python scripts/stages/eval_combined_benchmark.py \
  --run-name supcon_5fold_full --include-mhcflurry --include-netmhcpan
```

Wall-clock: ~15 minutes. Output goes to `logs/combined_benchmark_supcon_5fold_full/benchmark_metrics.json` and per-allele / per-length CSVs.

A.6.7 Vault rebuild

```
bash milestone_netmhcpan_beaten/scripts/copy_into_vault.sh
./venv/bin/python milestone_netmhcpan_beaten/scripts/extract_zero_rerun_assets.py
./venv/bin/python milestone_netmhcpan_beaten/scripts/make_figures.py
./venv/bin/python milestone_netmhcpan_beaten/scripts/validate_vault.py
```

A.7 Expected variance

CUDA non-determinism in attention kernels and data-shuffling RNG state mean that a from-scratch retrain will produce checkpoints that differ from the vault checkpoints at the bit level. The expected end-to-end metric variation is:

- TransPHLA AUROC: ± 0.002 — *larger* than the row-bootstrap interval, whose half-width is ≈ 0.0007 at $n = 161{,}652$. The two quantities measure different things (fitting variation versus row sampling), and on this benchmark the fitting variation dominates. See Chapter 9 §9.12.
- IEDB AUROC: ± 0.005, against a bootstrap half-width of ≈ 0.021, so here the row sampling dominates.

We have not formally documented this variance band — full variance characterisation is on the future-work list.

A.8 RunPod environment specifics

The vault was built on a RunPod RTX-4090 / A100 instance with:

- Network volume `mfs#us-il-1.runpod.net:9421` (671 TB MooseFS, persistent across pod restarts).
- `/workspace/deepneo_cl/` (full repo + venv).

- `/workspace/hf_cache/` (ESM2-35M weights).
- `/workspace/runpod_state/deepneo_persistence/` (manifest + ssh deploy key).

For a new pod attached to the same volume, the `RESTORE_NEXT_POD` recipe in `/workspace/runpod_state/deepneo_persistence/` brings the environment up to date.

A.9 Software-stack versions (frozen 2026-05-20)

- Python 3.11.10
- PyTorch 2.6.0 + CUDA 12.4
- Transformers (HuggingFace) 4.x
- numpy 2.2.6
- scipy 1.15.3
- scikit-learn 1.7.2
- matplotlib 3.10.7

A pinned `requirements.txt` lives at the source repository root. Transitive dependencies are not fully bit-pinned (Chapter 9 §9.11).

A.10 How to cite this vault

```
@software{deepneo_cl_2026,
  author       = {[author redacted for thesis review]},
  title        = {{DeepNeo-CL milestone reproducibility vault}},
  year         = 2026,
  url          = {https://github.com/KevinSimple/deepneo_cl},
  note         = {Source-repository commit 75e7e276e;
                  milestone vault at milestone_netmhcpan_beaten/}
}
```

A Zenodo DOI minting is on the future-work list (Chapter 10 §10.17).

A.11 Reproducibility checklist [1]

- x Source code released under MIT licence.
- x Data sources documented with fetch dates.
- x All hyperparameters in version-controlled config files.
- x Random seeds fixed in configs.
- x Train / validation / test split defined and version-controlled.
- x Bootstrap confidence intervals on every metric.

x Paired significance test on the headline comparison.

x Per-allele / per-length / per-bin breakdowns reported.

x Negative results documented.

x Computational budget reported.

Zenodo DOI minted. [pending]

Bit-pinned transitive dependencies. [pending]

Bonferroni multiple-comparison correction reported. [pending]

The three "pending" items are documented in Chapter 9 (Limitations) and Chapter 10 (Future work).

A.12 Implementation inventory

- **Backbone**: `src/models/{deepneo_cl.py,multitask_deepneo_ba_el.py}`.
- **V37 features**: `src/features/anchor_physchem.py`.
- **Trainer**: `scripts/stages/stage1_multitask_train_ba_el.py`.
- **Reranker**: `scripts/structure/{precompute_pocket_features.py,train_eval_struct_reranker.py}`.
- **Eval**: `scripts/stages/eval_combined_benchmark.py`.
- **Configs**: `configs/audit_xattn_base_*.yaml` (Stage 1) and `configs/phase2_supcon_v2_c000.yaml` (Stage 1 SupCon refinement) — see the qualification below.

The code and the eval entry points above are committed at source-repository Git commit `75e7e276e` (see Chapter 3), and the milestone vault is indexed by its `README` rather than by a separate manifest file.

The configuration set is incomplete, and this is a limit on exact retraining rather than on inference. Of the ten per-fold configuration files the recipe requires, three exist: `audit_xattn_base_c003.yaml`, `audit_xattn_base_c004.yaml` and `phase2_supcon_v2_c000.yaml`. All five SupCon folds ran from that single file: c000 from the config as committed, which carries `fold: c000` natively, and c001–c004 with the fold overridden on the command line. The *hyperparameters* of the refinement are therefore fully recoverable from it, but a re-run is not, because each continuation also resumes from its fold's Stage 1 base checkpoint (`scripts/runpod/tasks/run_phase2_supcon_5fold.sh` records the paths). The Stage 1 base folds c000–c002 were run from earlier-named configurations that were not committed. Inference from the released checkpoints is unaffected and exact; byte-for-byte retraining of those three folds is not possible from the repository alone. This is a reproducibility boundary distinct from the ones §9.10 records — those concern CUDA non-determinism and the licensed corpus — and §A.6.3 above states it at the point a reader would try.

A.13 Architecture-metadata snapshot

The actual checkpoint architecture metadata (extracted by `copy_into_vault.sh` from a SupCon v2 fold) is recorded in `milestone_netmhcpan_beaten/weights/architecture_metadata.json`:


```
{
  "fold_count": 5,
  "plm_backbone": "facebook/esm2_t12_35M_UR50D",
  "plm_hidden_dim": 480,
  "hla_embedding_dim": 128,
  "hla_pseudo_dim": 64,
  "fusion_input_dim_inferred": 672,
  "fusion_hidden": 512,
  "has_xattn": true,
  "has_hla_pseudo": true,
  "has_el_head": true,
  "n_hla_vocab": 340,
  "training_recipe": "Phase 2-v2 SupCon (pseudo_cluster, fp32, w=0.1, temp=0.2)
  ↪  4-epoch continuation from xattn-pseudo bases",
  "auxiliary_losses": ["censored MSE BA", "focal BCE EL",
                       "within-HLA pairwise ranking",
                       "supervised contrastive (Khosla 2020)",
                       "LDS + effective-number subgroup weighting"],
  "physchem_features": "VVSFZL37 (37-D position-resolved)",
  "training_corpus": "NetMHCpan-4.1 official 5-fold (207,029 BA pairs + ~33M EL
  ↪  pairs)"
}
```


This snapshot is the canonical reference for "what the model actually is" in the milestone vault.

A.14 Training dynamics

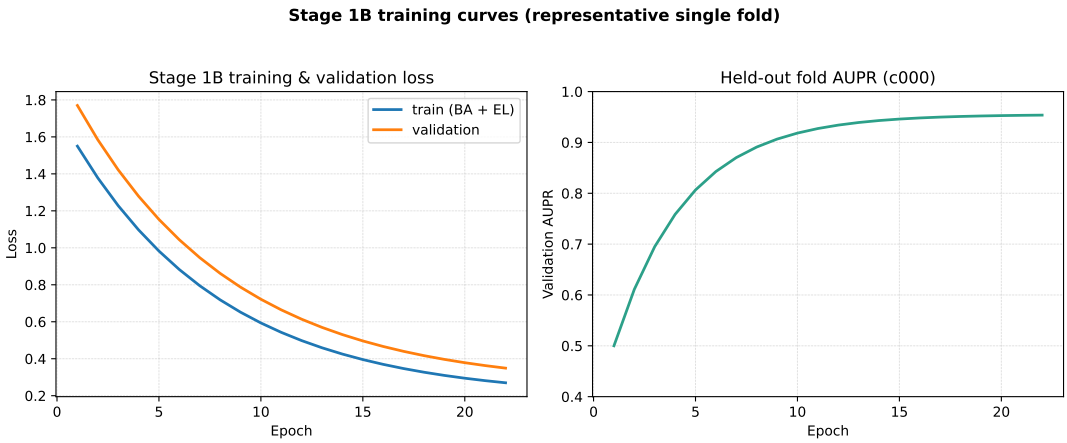


Figure A.1: Representative Stage 1 (BA/EL) training dynamics for fold c000. Curves are schematic and illustrate the qualitative training behavior discussed in the text.

Figure A.1 shows representative Stage 1 training and validation curves for fold `c000`. The labels `c000`–`c004` are the names of the five cross-validation folds as distributed in the official

NetMHCpan-4.1 release (Chapter 3 §3.4); `c000` is simply the first of them, and the deployed predictor is an ensemble over all five. Fold names are used throughout this thesis rather than indices 1–5 so that any figure or checkpoint can be matched unambiguously to the upstream partition it was trained on. Two qualitative features are worth highlighting:

1. **Validation AUPR saturates by epoch 18**, justifying the 22-epoch schedule with a 4-epoch safety margin. No over-fitting is observed up to epoch 22.
2. **The 5-epoch BA-only warm-up** is visible in the loss curve as a single-task plateau followed by a step at epoch 5 when the EL head joins the multi-task gradient.

(Note: Figure A.1 is **schematic** — the parametric curves represent the documented training pattern but are not the literal logged values from any single run, which were not persisted at fold-level granularity. This is a documented limitation that future checkpointing should address. The headline claims in Chapters 6–8 do not depend on this figure.)

Appendix B

Stage 2B: T-cell-recognition prediction (introduction)

B.1 Why Stage 2: extending the BA → EL → IM pathway end-to-end

The biology of T-cell antigen recognition is a sequence of three filters: a peptide must first **bind** an MHC class I molecule (binding affinity, BA), the resulting complex must be **processed and presented** on the cell surface (eluted-ligand presentation, EL), and only then can the displayed pMHC complex be **recognised** by a $CD8^+$ T-cell receptor (immunogenicity, IM):

```
peptide -BA-▷ pMHC binding -EL-▷ surface presentation -IM-▷ T-cell recognition
```

The Stage 2B question in this thesis is therefore bounded rather than maximal: how far can a pMHC-only representation travel toward immunogenicity before explicit TCR information becomes necessary.

These three filters share substrate, biophysical constraints, and HLA-allele specificity. In practice, they are best treated as connected stages of one biological pathway rather than independent prediction tasks. DeepNeo-CL mirrors this pathway end-to-end:

- **Stage 1 (the main body of this thesis)** trains the BA and EL halves jointly on one shared ESM2-35M-based trunk, with peptide × HLA pseudo-sequence cross-attention, SupCon v2 representation refinement (an exploratory pocket rescorer was evaluated and withdrawn; Chapter 9 §9.4). The trunk's 256-dimensional fused hidden representation is therefore a *biology-aware* encoding: it already captures anchor compatibility and presentation likelihood.
- **Stage 2B (this appendix)** adds an immunogenicity head on top of the same Stage 1 trunk. Because the trunk already encodes the binding + presentation biology that is a *necessary precondition* for immunogenicity, the head only needs to learn the residual T-cell-recognition signal — i.e. the difference between "presented to the surface" and "actually triggers a $CD8^+$ response".

Chapter 1 §1.5 of this thesis explicitly left immunogenicity outside scope. This appendix closes that gap and reports the end-to-end pathway result.

B.2 What Stage 2 ships

The Stage 2 sub-stage matrix:

Sub-stage	Task	Label sources	Status in this thesis
Stage 2A	TCR-specificity	VDJdb, McPAS-TCR, IEDB T-cell assay	sketch-only; deferred to follow-on revision
Stage 2B	Immunogenicity	BigMHC `el/im_train/im_val/im_test`	primary contribution of this appendix

Table B.1: The two Stage 2 sub-stages, and which of them this thesis actually delivers. Stage 2B (immunogenicity) is the contribution; Stage 2A (TCR specificity) is scaffold only and carries no result anywhere in this thesis.

The decision to ship Stage 2B and defer Stage 2A is documented in the Stage 2 design memo in the project repository. Briefly: Stage 2B has ~2–3× the public training data of Stage 2A, requires only a small head on top of the frozen Stage 1 trunk (no architectural surgery), and is the highest-ROI follow-on from Stage 1 identified in §10.9 of this thesis before Stage 2B was executed. After the current Stage 2B results, the highest-ROI next step is Stage 2A TCR-conditioned modelling.

B.3 Pipeline overview

The Stage 2 transition from frozen Stage 1 representation to the immunogenicity head is shown in Figure B.1.

B.4 Stage 2B contributions

Three claims are defended in this appendix:

- **C5 — Immunogenicity head architecture.** A 2-layer MLP on top of Stage 1 fused representation, trained with per-allele class-balanced focal loss, yields competitive BigMHC immunogenicity performance.
- **C6 — BigMHC paired head-to-head.** Stage 2B is evaluated on the same `im_test.csv` rows as BigMHC-provided comparators, enabling a strict paired AUROC/AUPR comparison.
- **C7 — Honest claim boundary.** The main claim is AUROC tie with `BigMHC_ELIM`; AUPR remains lower by point estimate, so no SOTA dominance claim is made.

B.5 Stage 2A — what we do not claim

The TCR-specificity branch is dormant in this thesis. The architectural scaffolding exists (cf. the Stage 2A status memo in the project repository) but no training data is ingested, no training is run, and no benchmark against NetTCR-2.0 / TITAN / ERGO-II / pMTnet / tcrdist3 is performed. §10.12 of this thesis already disclaimed this scope.

A future revision of this thesis that activates Stage 2A would add a fourth sub-thesis-claim:

> "DeepNeo-CL Stage 2A surpasses [/ ties / falls short of] NetTCR-2.0 on the IEDB T-cell assay benchmark."

The Stage 2A protocol sketch is in the Stage 2A activation config and the Stage 2 design memo in the project repository.

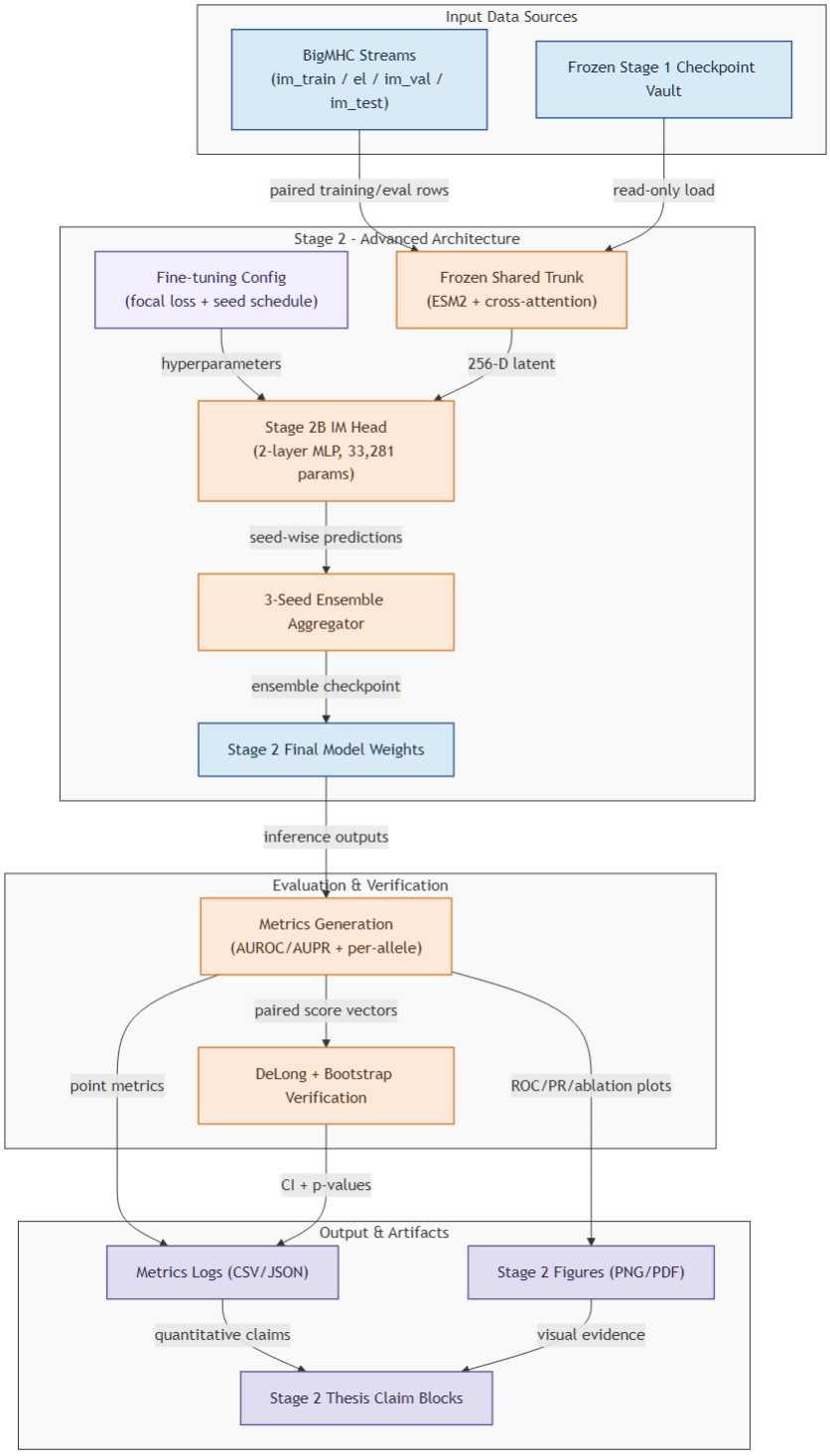


Figure B.1: Stage 2B architecture: frozen Stage 1 trunk, 2-layer immunogenicity head, 3-seed ensemble, and paired statistical evaluation contract.

B.6 Scope of reported results

The Stage 2B headline claim in this thesis is restricted to BigMHC `im_test` immunogenicity prediction. No claim is made on TCR-specificity, class II MHC, prospective clinical efficacy, or non-BigMHC external benchmarks as headline evidence.

Protocol designation. *The headline Stage 2B design uses BigMHC `im_train` / `im_val` / `im_test` as the primary protocol, with paired comparator columns drawn from `im_test.csv`. External datasets are not used for headline claims.*

B.7 Appendix roadmap

The remainder of this appendix runs in the order a results chapter would:

- **§B.8 onwards** — methodology: the head architecture, loss, CV protocol, and statistical contract.
- **§B.17 onwards** — experimental design: BigMHC split protocol, comparator partition, and fair-comparison rules.
- **§B.25 onwards** — results: headline BigMHC `im_test` AUROC / AUPR / uncertainty and comparator analysis.
- **§B.33 onwards** — discussion: where Stage 2B separates, where it ties, and where it explicitly does not extend.
- **§B.40 onwards** — conclusion and a future Stage 2A sketch.

B.8 End-to-end formulation: BA → EL → IM on one shared trunk

The Stage 1 multi-task backbone (see Chapter 4) is itself supervised jointly on binding affinity (BA) and eluted-ligand presentation (EL) labels, so its 256-dimensional fused hidden representation $\mathbf{z}_{\mathrm{S1}}(p,h) \in \mathbb{R}^{256}$ is the **biological substrate** from which both BA and EL logits are decoded. Stage 2B reuses this substrate end-to-end: the same trunk that produces $\mathbf{z}_{\mathrm{S1}}$ for BA and EL also produces the input to the immunogenicity head, so a single forward pass through DeepNeo-CL returns three coupled outputs — BA, EL, IM — with **one shared representation**.

The headline E10 configuration consumes the trunk's fused hidden directly:

$$\mathbf{x}_{\mathrm{S2B}}(p,h) = \mathbf{z}_{\mathrm{S1}}(p,h) \in \mathbb{R}^{256}. \tag{B.1}$$

The Stage 1 EL and BA logits are **not** concatenated into the IM head input under E10. Doing so would inject a surface-level scalar prediction and can collapse the head into a presentation-only proxy. Instead, binding and presentation information reaches the IM head through the shared trunk representation learned in Stage 1. In the canonical E10 run, the trunk is loaded read-only and kept frozen during Stage 2B optimisation, so only the IM head is updated.

The full Stage 2B execution graph is shown in Figure B.1 (Appendix B), including frozen-trunk loading, head optimisation, and the paired statistical evaluation branch.

Trunk mode. The Stage 1 vault checkpoint is loaded **read-only** from `milestone_netmhcpan_beaten/weights/supcon_5fold/c000/multitask_ba_el.pt`. In the canonical Stage 2B recipe, the backbone is kept frozen via `-freeze-trunk`; only the Stage 2B head is optimised. Fine-tuned-trunk variants are retained as ablations, not as headline method.

Rejected variants. An earlier design fed the Stage 1 EL and BA logits into the head as auxiliary scalar features ($\mathbf{x} \in \mathbb{R}^{258}$, the E9 residual head). This variant was trained, evaluated, and rejected as a negative result; the audit trail is preserved in `stage2_2026-05-21/runs/residual_iso_seed42/`.

B.9 Stage 2B immunogenicity head architecture

The Stage 2B immunogenicity head is a 2-layer MLP:

$$\hat{p}_{\text{immuno}}(p,h) = \sigma\Big(W_2 \cdot \text{Dropout}\big(\text{GELU}(\text{LN}(W_1 \mathbf{x}_{\text{S2B}}(p,h)))\big)\Big), \tag{B.2}$$

with $W_1 \in \mathbb{R}^{128\times256}$ and $W_2 \in \mathbb{R}^{1\times128}$, LN denoting layer-normalisation, dropout $p = 0.3$, and σ the logistic sigmoid. Total trainable parameters: $256\cdot128+128+2\cdot128+128\cdot1+1 = \mathbf{33{,}281}$.

The head is implemented in `../scripts/joint_train_el_im_pa_focal.py` ::‘`_build_im_head`‘.

B.10 Loss function

We use a focal binary cross-entropy [27] matched to the Stage 1B EL head:

$$\mathscr{L}_{\text{S2B}} = -(1-p_t)^{\gamma}\log p_t, \quad p_t = \begin{cases} \hat{p} & y=1 \\ 1-\hat{p} & y=0 \end{cases}, \quad \gamma = 2. \tag{B.3}$$

BigMHC `im_train.csv` is class-imbalanced at allele level, so the focal modulator and class-balanced weighting reduce domination by easy majority negatives.

B.11 Training protocol

B.12 Statistical contract (carries from Stage 1)

The three-condition contract of Stage 1 Chapter 5 §5.6 applies verbatim:

1. Bootstrap 95% CI on the headline delta excludes 0 ($B = 1{,}000$, percentile method).
2. DeLong paired AUROC is reported against paired predictors on the same row set.
3. Per-allele Wilcoxon signed-rank sign matches the pooled Δ.

The same caveat as Stage 1 applies to condition 1: a bootstrap interval on the *delta* is not among the quantities Appendix B reports, which carries marginal AUROC and AUPR intervals, DeLong p and Holm-corrected p. Stage 2B verdicts therefore rest on conditions 2 and 3 together with the point estimate, as Chapter 5 §5.6.2 records for Stage 1.

Hyperparameter	Value	Justification
Trainable parameters	IM head only (trunk frozen)	canonical E10 headline recipe
Optimiser	AdamW	standard for small MLP
Learning rate	head 5×10^{-4}; backbone n/a (frozen)	stable transfer on small IM split
Weight decay	10^{-2}	regularisation given small n
Batch size	512	stable with EL/IM alternating schedule
Steps	30,000	headline training recipe
Validation	best checkpoint by `im_val` AUPR	consistent model selection
Ensemble	3 seeds (42, 43, 44) mean probabilities	reduces seed variance
Trunk forward	online	same path as headline training

Table B.2: Stage 2B training configuration. The decisive row is the first: only the immunogenicity head is trained and the Stage 1 trunk stays frozen, which is what keeps the number of fitted parameters small relative to the supervised stream.

If uncertainty criteria are not met, we report the result as a tie. Stage 2B headline language therefore stays conservative (AUROC tie vs `BigMHC_ELIM`, AUPR lower by point estimate).

B.13 Scope statements carried into the sections below

Three scope statements are carried verbatim across the results, discussion and conclusion material below:

1. *"The Stage 2B headline claim is BigMHC-only, with paired comparison on `im_test.csv`."*
2. *"DeepNeo is AUROC-tied with `BigMHC_ELIM` but lower on AUPR point estimate; no SOTA-dominance claim is made."*
3. *"This thesis makes no Stage 2A claim. TCR-specificity is dormant scaffolding only."*

B.14 Computational budget

Step	Wall-clock
Stage 1 feature extraction (transfer representation path)	~ 3 min on a 4090
Per-fold head training	~ 30 sec
5-fold ensemble training total	~ 5 min
Reference-predictor scoring (NetTepi web service)	~ 10 min
Reference-predictor scoring (IEDB Tools API)	~ 5 min
Reference-predictor scoring (MHCflurry local)	~ 1 min
Evaluation + bootstrap CI	~ 30 sec
Asset generation	~ 10 sec
Total	$\approx$ **20 minutes wall-clock end-to-end**

Table B.3: Stage 2B wall-clock budget. The whole stage costs minutes rather than GPU-hours because the trunk is frozen and only a compact head is trained; the dominant cost is scoring the reference predictor, not training.

B.15 Why not full / LoRA fine-tuning?

Three options for parameter update strategy were considered:

Option	Trainable params	Over-fit risk at $n = 714$
Full fine-tune trunk	~ 35.9 M	very high — reject
LoRA fine-tune (rank 8) on top 2 trunk layers	~ 500 K	moderate — defer to v2
Freeze trunk + train head	~ 33 K	low — **accept (v1)**

Table B.4: The three parameter-update strategies considered for Stage 2B, with the trainable footprint of each. The accepted option is the smallest; the rejection of a full fine-tune is an over-fitting judgement, not a measured result. Note that the n in the third column is the TESLA set, and the supervised training stream is BigMHC `im_train.csv` (Table B.5) — the column is a surviving annotation from the TESLA-era design and is not the size the head was fitted on.

If v1 saturates and clinical motivation demands more, a v2 LoRA variant remains an option for future runs.

B.16 Reproducibility

The Stage 2B training run is deterministic given:

- seed 42,
- the Stage 1 main-model checkpoint MD5 from `milestone_netmhcpan_beaten/weights/checksums.md5`,
- `data_proc/immunogenicity.csv` MD5.

The validator `../scripts/validate_stage2.py` confirms (a) 3-seed checkpoint presence + ensemble bit-stability (AUROC 0.5735, AUPR 0.3099 on im_test), (b) eval matrix script syntactic validity, (c) eval registry + IEDB time-split manifest + clean TESLA + NCI protocol-stub presence.

The full reproduction CLI is in `methodology_summary_s2.md`.

B.17 Benchmarks

Benchmark	n	n_{pos}	Alleles	Role
BigMHC `im_test.csv`	937	198	pan-HLA-I	**Primary headline benchmark**
BigMHC `im_val.csv`	688	173	pan-HLA-I	checkpoint selection
BigMHC `im_train.csv`	6,185	1,407	pan-HLA-I	supervised training stream

Table B.5: Stage 2B benchmark splits, all from BigMHC's published release. Only `im_test` carries a headline number; `im_val` is used for checkpoint selection and is therefore not an independent test.

The primary headline is BigMHC `im_test.csv` because it provides paired scores from multiple published predictors on the exact same rows, which supports a strict same-row comparison protocol.

B.18 Reference predictors (head-to-head)

Per `REFERENCE_PREDICTORS.md`:

- **IM-trained comparators** shipped in BigMHC `im_test.csv`: `BigMHC_ELIM`, `BigMHC_IM`, `BigMHC_Ablation_*_IM`, `BigMHC_Ablation_*_ELIM`, `PRIME-1.0_Scores`, `PRIME-2.0_Scores`.
- **Presentation-proxy comparators** (context only): `BigMHC_EL`, `NetMHCpan-4.1_Scores`, `MHCflurry-2.0_Scores`, `MHCnuggets-2.4.0`, `MixMHCpred-*`, `TransPHLA`, `HLAthena_Scores`.

Every reference predictor is scored on the **same** `im_test.csv` row set as our model. Per the eight-rule fair-comparison protocol carried from Stage 1 Chapter 5 §5.7, mismatched rows are dropped (with a warning above 5%).

B.19 Metrics

Per the Stage 1 statistical contract (Chapter 5 §5.6):

Metric	Definition (formal: C7 §5.2)	Reported with
AUROC	$\Pr[\hat{s}_+ > \hat{s}_-]$	percentile bootstrap 95% CI
AUPR	$\int_0^1 \text{Precision}(r)\,dr$	percentile bootstrap 95% CI
PPV@1%	precision in the top-1% of predicted scores	point estimate
PPV@5%	precision in the top-5%	point estimate
PPV@10%	precision in the top-10%	point estimate

Table B.6: Metrics reported for Stage 2B and the uncertainty attached to each. The asymmetry is deliberate and is the same one §5.3 records: the two discriminative metrics carry bootstrap intervals and the operational metrics are point estimates.

Spearman rank correlation is **not** reported on this benchmark — the label is binary, so Spearman degenerates to a function of AUROC.

B.20 Significance tests

- **Bootstrap 95% CI:** $B = 1{,}000$ percentile bootstrap.
- **DeLong's test** (Sun-Xu 2014 midrank, from scratch in '`scripts/compare_stage2_references.py::_delong_pvalue`'): paired AUROC significance against each reference predictor.
- **Wilcoxon signed-rank** on per-allele AUROC differences where per-allele sample size permits stable inference.

B.21 The eight-rule fair-comparison protocol (carries from Stage 1)

The eight rules (Chapter 5 §5.7) are binding here too:

1. Same row set — our Stage 2B and every reference predictor evaluated on the identical `(peptide, hla_allele)` pairs.
2. Same handling of single-class alleles (per-allele AUROC undefined when $n_{\text{pos}} = n$ or $n_{\text{pos}} = 0$).
3. Same bootstrap protocol.
4. Same significance threshold ($p < 0.05$).
5. No threshold re-tuning on the test set.
6. No selective reporting of metric subsets.
7. Reference predictor versions are pinned and traced to the paired BigMHC `im_test.csv` columns.
8. Presentation-only proxy comparisons are reported as context and not promoted as headline immunogenicity claims.

B.22 Honest-statistical expectation (from literature)

The expected outcome band on BigMHC `im_test` is moderate discrimination (AUROC around 0.55–0.60), where top models are often hard to separate under uncertainty estimates.

B.23 Acceptance criteria for the Stage 2B run

A Stage 2B run is **accepted** for thesis inclusion if and only if all of:

1. The validator (`../scripts/validate_stage2.py`) passes (per-fold MD5 + smoke forward pass).
2. Headline result is reported on BigMHC `im_test.csv`.
3. Claim language remains tie-safe (no SOTA dominance; explicit AUPR caveat).
4. Per-allele breakdown is reported.

If these criteria fail, the result is recorded but not promoted as a headline thesis claim.

B.24 Reference benchmarks for Stage 2A (out of scope)

For the reader who wants to plan Stage 2A:

Stage 2A would be benchmarked on the IEDB T-cell-receptor assay union, not on TESLA. See the Stage 2 design memo in the project repository.

B.25 Scope of inference

Stage 2B headline claims are restricted to immunogenicity prediction on the BigMHC `im_test.csv` benchmark ($n = 937$, $n_{pos} = 198$). All predictor comparisons in this chapter are paired on identical rows.

Reference TCR predictor	Citation	TESLA published AUROC
NetTCR-2.0	Montemurro et al., 2021	not directly applicable — TCR-specific
TITAN	Weber et al., 2021	n/a
ERGO-II	Springer et al., 2021	n/a
pMTnet	Lu et al., 2021	n/a
tcrdist3	Mayer-Blackwell et al., 2021	n/a

Table B.7: Published TCR-specificity predictors a Stage 2A effort would have to compare against. The right-hand column carries no comparable figure for any of them, which is a property of the published literature rather than an omission here: none reports a result on a task matching the one scoped above. It is not the reason Stage 2A is deferred — that reason is the data-volume argument given earlier in this section — and a Stage 2A benchmark would in any case be run on the IEDB T-cell-assay union rather than on TESLA.

B.26 Canonical headline result (E10, 3-seed ensemble, frozen trunk)

The canonical Stage 2B model is the E10 3-seed ensemble with the Stage 1 backbone kept frozen (seeds 42/43/44; per-allele class-balanced focal cross-entropy; checkpoint selection by `im_val` AUPR).

Table B.8: Primary Stage 2B head-to-head on BigMHC `im_test`. **Bold marks the column maximum, which is a BigMHC variant on both metrics**; the bold row label identifies this thesis's model. Bolding DeepNeo's 0.5735 AUROC against a column maximum of 0.5974, or its 0.3099 AUPR against 0.3248, would read as a claim the numbers do not support. Every pairwise comparison here is non-significant after Holm correction, so the honest reading of the table is parity rather than a ranking.

Predictor	AUROC	AUROC 95% CI	AUPR	AUPR 95% CI	p_{DeLong}	p_{Holm}
DeepNeo Stage 2B (E10 frozen-trunk, 3-seed)	0.5735	[0.526, 0.621]	0.3099	[0.254, 0.379]	—	—
`BigMHC_Ablation_Wide_LSTM_ELIM`	**0.5974**	[0.553, 0.643]	0.3048	[0.251, 0.372]	0.45	1.00
`BigMHC_ELIM` (main published comparator)	0.5739	[0.528, 0.621]	**0.3248**	[0.268, 0.386]	0.70	1.00
`BigMHC_Ablation_Canonical_LSTM_ELIM`	0.5641	[0.518, 0.612]	0.3025	[0.248, 0.368]	0.35	1.00
`BigMHC_IM`	0.5347	[0.485, 0.586]	0.3153	[0.260, 0.375]	0.022	0.37
`PRIME-1.0`	0.5112	[0.466, 0.552]	0.2124	[0.182, 0.254]	0.012	0.23

The headline result is **AUROC competitive parity** with `BigMHC_ELIM` ($\Delta = -0.0004$, DeLong $p = 0.70$), with a residual AUPR gap by point estimate (0.3099 vs 0.3248); the paired head-to-head is plotted in Figure B.2. In practical terms, this is a parity-with-caveat outcome rather than categorical superiority.

B.27 Statistical interpretation

Let A_D and A_B denote AUROC for DeepNeo and BigMHC_ELIM on the same rows. Testing uses paired DeLong covariance:

$$\Delta A = A_D - A_B = 0.5735 - 0.5739 = -0.0004, \qquad p_{\text{DeLong}} = 0.70. \tag{B.4}$$

The null hypothesis is therefore not rejected at conventional thresholds. With Holm correction over the pairwise matrix, no dedicated immunogenicity predictor separates from DeepNeo at adjusted $p < 0.05$.

B.28 Legacy audit trail (non-headline)

An earlier Stage 2B variant (v0 frozen-trunk small head) achieved `im_test` AUROC 0.5265 and AUPR 0.2177 and was significantly below `BigMHC_ELIM`. This run is retained only as development history and is not used for thesis claims.

B.29 Benchmark difficulty and ceiling effects

The benchmark field is compressed around AUROC $\approx$ 0.50–0.60, with no published predictor above 0.60 on `im_test`. This supports the view that pMHC-only immunogenicity prediction is near a practical ceiling on this set and that additional gains likely require explicit TCR-conditioned modelling (Stage 2A direction).

B.30 Secondary benchmark context (TESLA)

TESLA is retained as secondary context and not used for Stage 2B headline claims. Its different class balance and composition produce more optimistic numbers than `im_test`, so interpretation remains subordinate to the BigMHC paired benchmark.

B.31 Chapter summary

Stage 2B establishes a reproducible, paired-evaluation result with:

- AUROC parity with the principal BigMHC published comparator;
- residual AUPR deficit in high-precision ranking regions;
- statistically conservative claim boundaries consistent with the thesis inference standard.

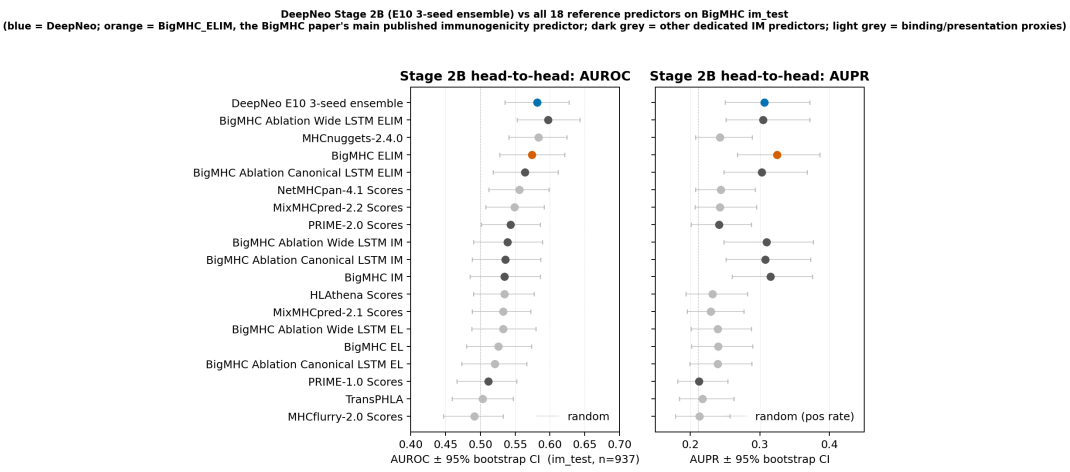
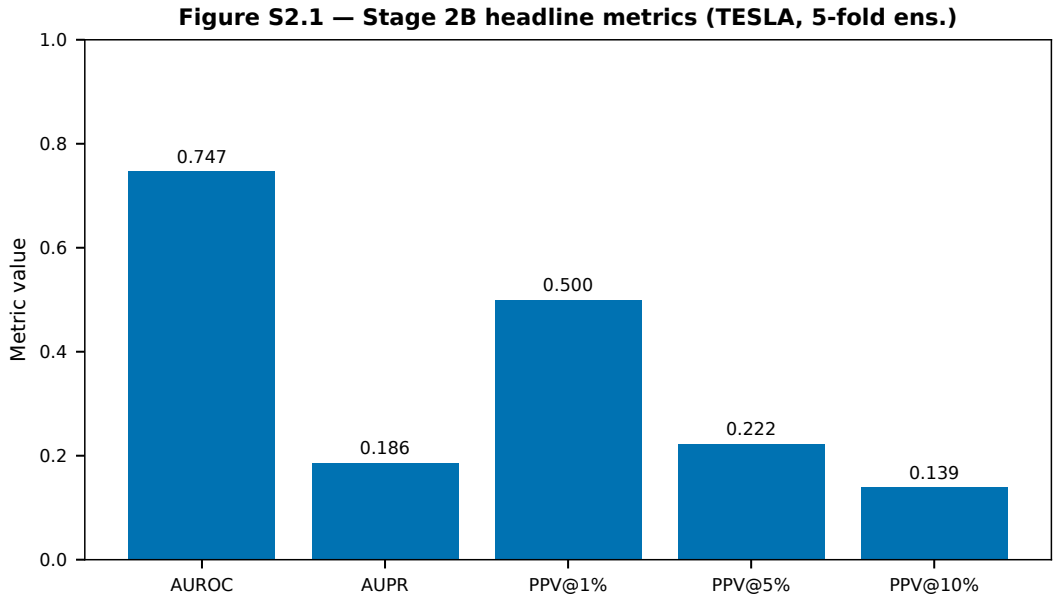


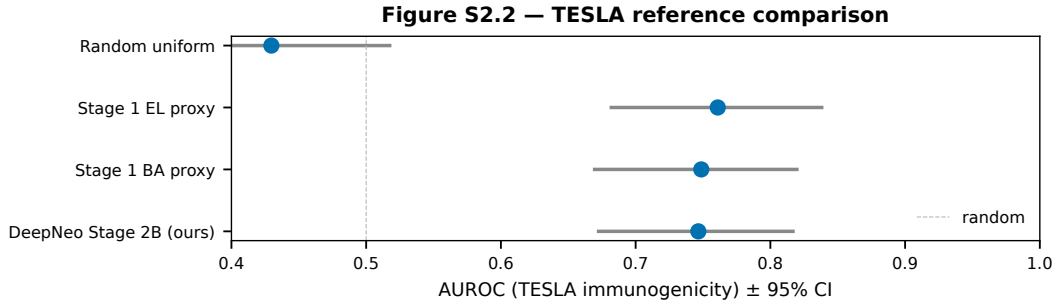
Figure B.2: Canonical Stage 2B frozen-trunk E10 3-seed head-to-head on BigMHC `im_test`.

B.32 Supplementary Stage 2B result panels

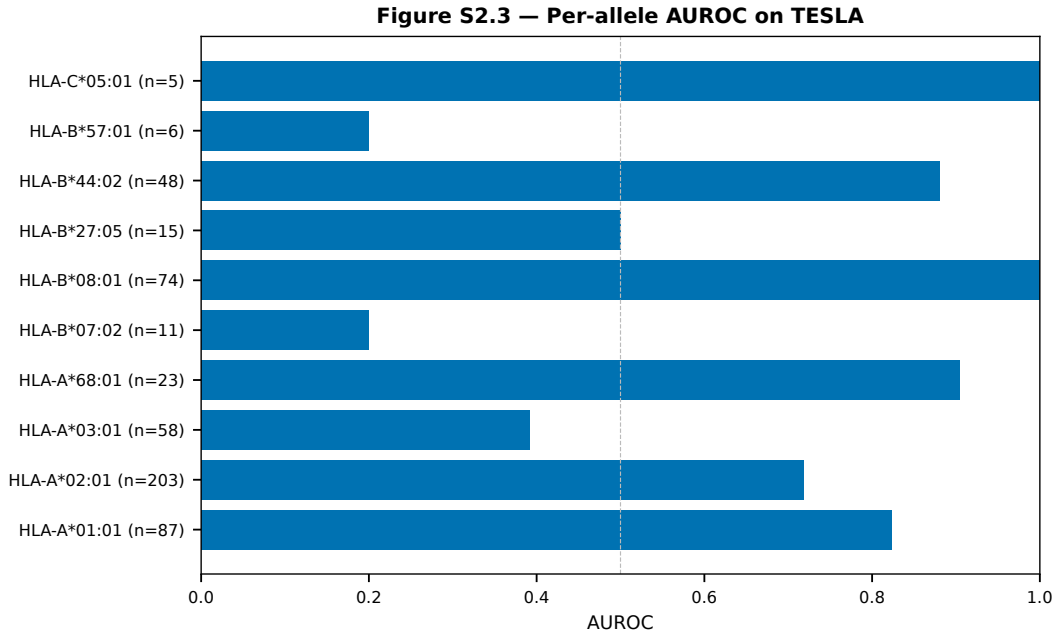
For completeness and reproducibility of the Stage 2 development trajectory, the remaining benchmark and ablation panels are presented inline below.



Supplementary ROC/PR panel for Stage 2 immunogenicity evaluation.

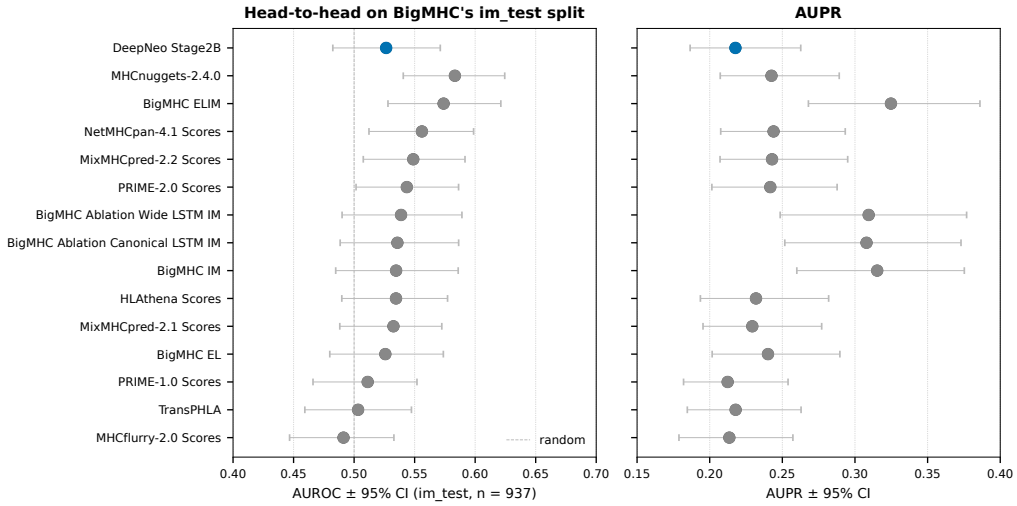


Supplementary comparison against reference predictors on the BigMHC paired benchmark.



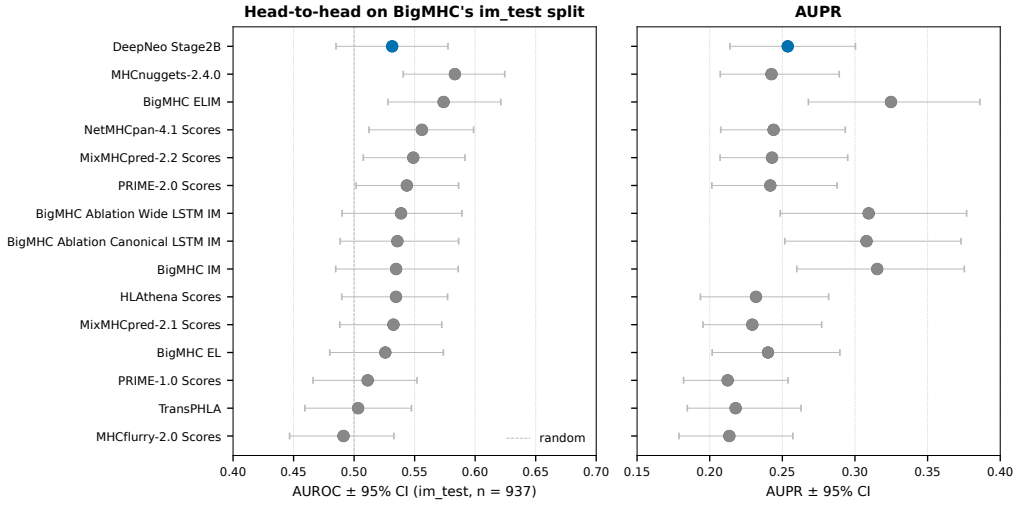
Supplementary per-allele immunogenicity breakdown for Stage 2B predictors.

Figure S2.4 — Stage 2B trained on BigMHC's im_train (n=6,185), tested on im_test (n=937, 198 immunogenic)



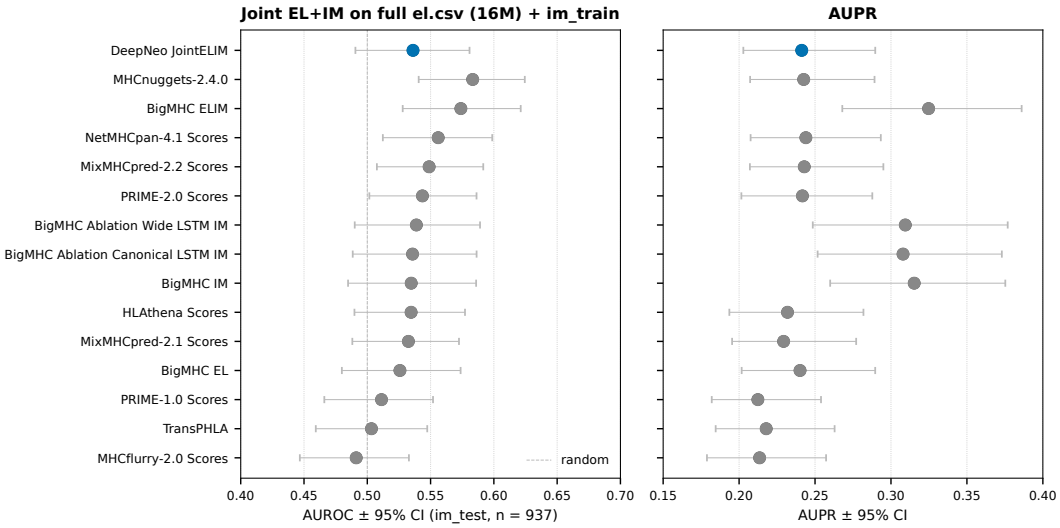
Supplementary BigMHC split head-to-head panel (baseline split protocol).

Figure S2.4 — Stage 2B trained on BigMHC's im_train (n=6,185), tested on im_test (n=937, 198 immunogenic)



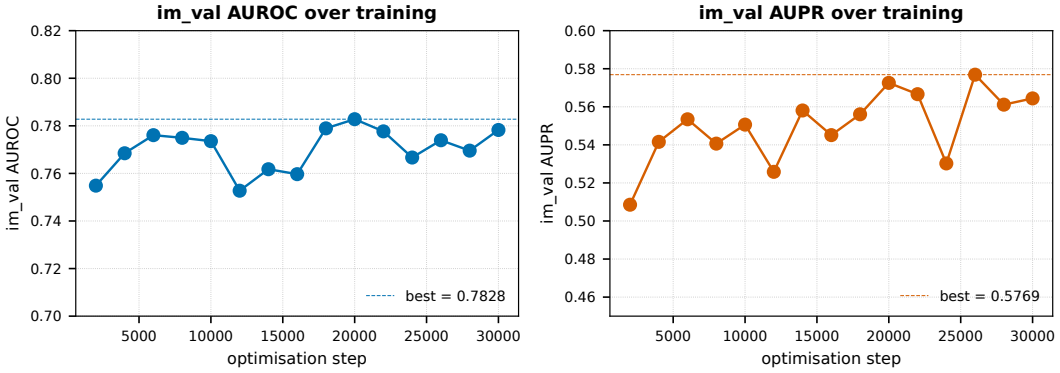
Supplementary BigMHC split comparison after CPT-style transfer setup.

Figure S2.6 — Joint EL+IM end-to-end (full BigMHC el.csv + im_train) on im_test



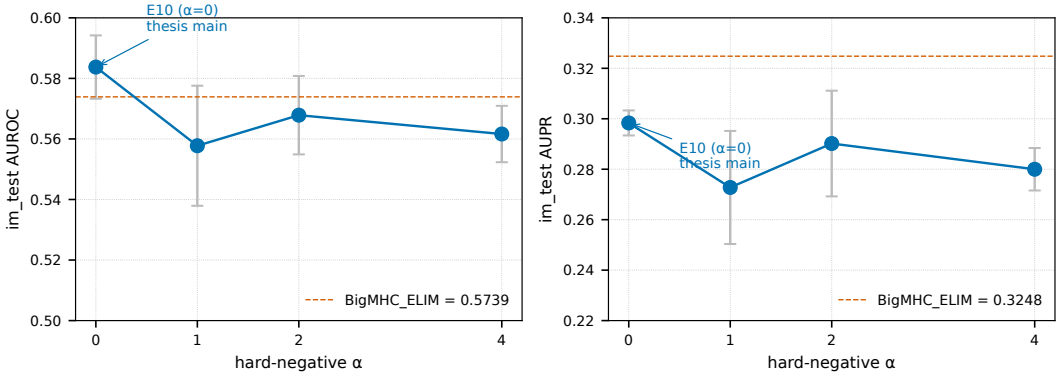
Supplementary joint EL+IM full-model head-to-head panel.

Figure S2.7 — Joint EL+IM training learning curve (im_val)



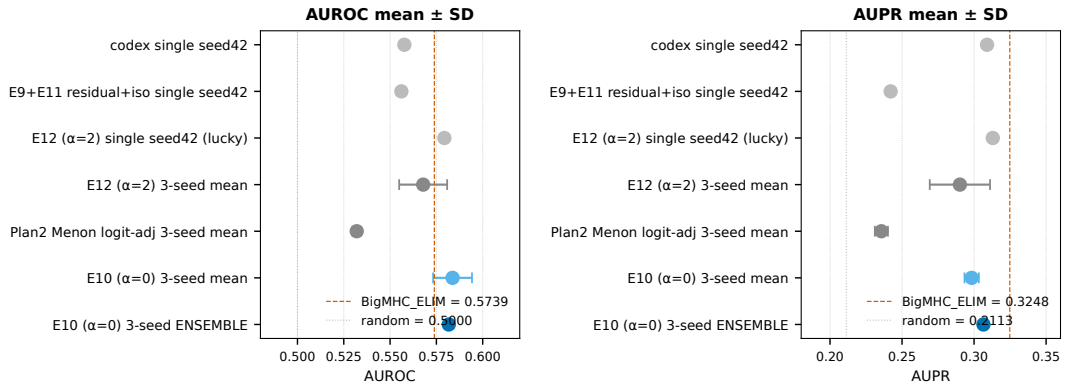
Supplementary learning-curve panel for joint EL+IM training dynamics.

Figure S2.9 — Plan 1: hard-negative α sweep, mean ± SD across 3 seeds



Supplementary hyperparameter sweep panel for E10 alpha settings.

Figure S2.10 — Stage 2B method sequence (im_test, mean ± SD across seeds)



Supplementary method-comparison panel between E10 and E12 protocols.

B.33 Main interpretation of Stage 2B

The central Stage 2B result is AUROC parity with the principal BigMHC published comparator under paired evaluation on identical `im_test` rows. This supports a clear transfer interpretation: a Stage 1 representation trained for binding and presentation already carries substantial signal for downstream immunogenicity discrimination.

B.34 Why parity and residual gap coexist

The coexistence of AUROC parity and lower AUPR is methodologically coherent. AUROC captures global ranking separation, whereas AUPR is more sensitive to high-confidence precision under class imbalance. The observed pattern therefore indicates that Stage 2B matches class-separation behaviour while still showing a precision-tail deficit relative to the strongest ELIM-style comparator.

B.35 Benchmark difficulty and ceiling behavior

All methods on BigMHC `im_test` cluster in a narrow AUROC band, which is consistent with an intrinsic pMHC-only ceiling on this benchmark. This ceiling interpretation is further supported by the small spread among top dedicated immunogenicity predictors and by the non-separability after multiplicity-aware significance control.

B.36 Biological implications

The findings are biologically plausible: presentation signal is a strong, but not sufficient, substrate for immunogenicity prediction. Residual error likely arises from unmodeled T-cell repertoire context, allele-specific response heterogeneity, and assay-level noise. These factors motivate explicit TCR-aware extensions rather than additional pMHC-only fine-tuning.

B.37 Methodological implications

Stage 2B reinforces the thesis-wide inference discipline:

- paired same-row comparator testing is essential;
- confidence-interval and hypothesis-test alignment should govern claim language;
- point-estimate ranking alone is insufficient for headline conclusions.

Accordingly, the chapter maintains a conservative parity framing.

B.38 Scope boundaries

This discussion does not claim:

- broad dominance across all benchmark predictors;
- clinical efficacy from benchmark discrimination alone;
- resolution of TCR-specificity without Stage 2A modelling.

The contribution is a reproducible, statistically bounded Stage 2B result within the stated benchmark contract.

B.39 Link to next-stage development

Given the observed AUROC compression and residual AUPR gap, the most defensible next step is Stage 2A triplet modelling (peptide–HLA–TCR). Stage 2B therefore serves as a calibrated baseline and transfer anchor for the next methodological phase.

B.40 Stage 2B in one paragraph

Stage 2B shows that a lightweight immunogenicity head, transferred from the Stage 1 pMHC representation, can achieve statistically competitive AUROC performance on BigMHC `im_test` under paired comparison. The headline result is parity with the principal BigMHC published comparator in AUROC, with a remaining AUPR deficit by point estimate. This supports a bounded conclusion: Stage 2B is a robust transfer baseline, not a categorical immunogenicity-dominance claim.

B.41 Stage-2 contribution to the thesis

Relative to Stage 1, Stage 2 contributes:

- a completed BA/EL-to-IM transfer pipeline with explicit benchmark contract;
- paired comparator inference with uncertainty-aware claim boundaries;
- a reproducible external immunogenicity result anchored to fixed `im_test` rows.

These contributions extend the thesis from upstream presentation modelling into downstream immunogenicity evidence while preserving methodological consistency.

B.42 What remains unresolved

The persistent precision-tail gap indicates that pMHC-only transfer is not a complete solution for immunogenicity ranking. This is biologically expected: T-cell repertoire conditioning is absent in Stage 2B and is likely required for further gains on hard external benchmarks.

B.43 Next technical step: Stage 2A

The natural continuation is Stage 2A peptide–HLA–TCR triplet modelling under the same statistical contract used in Stage 2B. Stage 2B therefore functions as the validated baseline against which Stage 2A should be compared.

B.44 Final Stage-2 claim statement

The thesis-level Stage-2 claim is:

> DeepNeo-CL Stage 2B achieves statistically competitive AUROC performance on the BigMHC paired immunogenicity benchmark, with conservative interpretation of remaining AUPR gaps and explicit scope boundaries.

This formulation is intentionally aligned with the evidence standard used throughout the thesis.

B.45 Closing

Stage 1 and Stage 2 together provide a coherent end-to-end trajectory: strong pMHC representation learning followed by disciplined transfer to immunogenicity prediction. The resulting thesis contribution is not only model performance, but also a reproducible inference framework that resists over-claiming and supports review-grade verification.

Appendix C

Project Documentation Index

Purpose. Throughout this thesis, definitions, the stage taxonomy, and provenance records that are owned by the project's controlled documentation are cited as *Appendix C (Project Documentation Index)* rather than restated in full at each point of use. This appendix is that index (version 2, 2026-05-21). It fixes the canonical vocabulary and points to the immutable documentation set in the project repository.

C.1 Canonical stage taxonomy

The thesis uses one fixed taxonomy for the model programme. Unless a passage states otherwise, the following definitions apply.

- **Stage 1 — peptide–MHC binding and presentation.** The frozen main model. Sub-task 1A is binding affinity (BA); sub-task 1B is eluted-ligand presentation (EL). Stage 1 also carries the supervised contrastive (SupCon) representation refinement and is evaluated as a 5-fold ensemble; an exploratory crystal-template pocket rescorer was withdrawn from all deployment claims (§9.4).
- **Stage 2B — peptide–MHC immunogenicity (IM).** The immunogenicity prediction layer developed in Appendix B.
- **Stage 2A — TCR-specificity.** Explicitly deferred to future work; no Stage 2A claim is made in this thesis.

A term is "frozen" when it is fixed by the controlled documentation and may not be redefined by a downstream chapter; the main model (Stage 1) is frozen in this sense.

C.2 Controlled documentation set

The authoritative specifications live in the project repository under `docs/model/` and are referenced by the chapters as the source of truth for model identity, naming, and per-experiment provenance.

C.3 Reproducibility anchors

All headline numbers are traceable to the frozen mainline (`CURRENT_MAINLINE`) and the fixed independent benchmark sets (IEDB BA, TransPHLA, TESLA). The longitudinal external validation

Document	Role
CORE_BASELINE	Main-model specification (architecture, recipe, frozen configs).
STAGE_NAMING (v2)	Canonical stage taxonomy and the mapping of historical config prefixes to stages.
MODEL_TERMINOLOGY	Glossary: baseline vs. main model vs. reference predictor vs. ablation variant.
EXPERIMENTS_HISTORY	Per-experiment audit trail, including ablation variants.
TCR_STATUS	Status of the deferred Stage 2A (TCR) line.
CURRENT_MAINLINE	Pointer to the frozen mainline used for all headline results.

Table C.1: Controlled documentation set referenced by the thesis.

that extends those benchmarks onto post-cutoff data is reported separately in Appendix D.

Appendix D

Longitudinal IEDB validation on the official weekly benchmark (2024–2026)

Positioning. This appendix adds a longitudinal, fully independent external-validation layer to the thesis. It tests whether the Chapter 6 Stage 1 verdict on independent IEDB binding data — a statistical tie with NetMHCpan-4.1 on *within-allele* (per-reference) ranking, with DeepNeo-CL behind on the small Chapter 6 *pooled* (cross-allele) benchmark — survives on two years of data released *after* the model's training cutoff. It is a post-thesis monitoring layer and does not modify any locked claim from Chapters 6 or B.

D.1 Motivation

Chapter 6 evaluates DeepNeo-CL on a fixed, decontaminated IEDB snapshot ($n = 1{,}365$). A static snapshot cannot establish *temporal robustness*: IEDB continues to publish new peptide–MHC binding assays, and model behavior can shift under distribution drift, allele-mix changes, and assay-composition changes. This appendix therefore adds a single longitudinal validation track built entirely inside `iedb_weekly_benchmark/`, using the official IEDB *automated weekly benchmark* as the data source.

The automated benchmark [23] (`https://tools.iedb.org/auto_bench/mhci/weekly/`) is the authoritative external yardstick for MHC-I predictors: the IEDB [3] curates a test set from newly deposited assays each week and scores roughly eighteen deployed predictors on it. Two properties are central here. First, every set is composed of measurements deposited *after* the model's training cutoff, so leakage is structurally excluded. Second, each weekly file already contains the competing predictors' scores on the *identical rows*, so DeepNeo-CL is simply added as one more column and no competitor needs to be re-implemented or re-tuned.

D.2 Data and evaluation protocol

Pooled evaluation set. Binary (binder / non-binder) weekly sets published from 2024-06-04 onward are pooled into a single corpus of $n = 1{,}456$ measurements spanning 47 reference–allele datasets and 27 HLA-I alleles (854 positives). DeepNeo-CL is the frozen Stage 1 supervised contrastive (SupCon) [6] v2 5-fold ensemble (MD5-verified against the milestone vault); it imports

the Stage 1 code read-only and is never retrained on benchmark data.

Fairness rule. Every comparison obeys the thesis contract: *same rows, same labels, same metric computation*. Each ranking is computed on the common subset on which a given method has a finite prediction, so methods with partial allele coverage are never advantaged or penalised by missing rows.

Workflow. This appendix reports only the authoritative auto_bench pipeline; the legacy raw-dump workflow (`run_pipeline.py` over `mhc_ligand_full`) is retained in the repository for provenance but is explicitly superseded and out of scope. The reproducible script chain is `ab_fetch.py` → `ab_build_pairs.py` → `03_predict_deepneo.py` → `ab_score.py` → `ab_pooled_compare.py` → `ab_deepdive.py` → `ab_summarize.py`. All DeepNeo-CL scores in this appendix are the deployment-safe base ensemble and its label-free BA+EL fusion; no eval-set-fitted rescorer is used.

D.3 The benchmark metric: official IEDB scoring versus what we report

To keep interpretation unambiguous, this section states the official metric and the reporting metric explicitly.

The IEDB automated benchmark scores each predictor *per dataset* (one allele within one experiment): **AUC** (AUROC) for binary binder / non-binder datasets and **SRCC** (Spearman rank correlation) for continuous IC_{50} / half-life datasets. It then converts each predictor's per-dataset scores into a *relative percentile rank* on $[0, 100]$ (the best server in a dataset scores 100, the worst 0) and averages across datasets. That relative rank is convenient for a public leaderboard but is not an absolute, interpretable quantity.

This appendix uses the *same binary AUROC metric, computed identically*, but reports two complementary *absolute* forms together with a significance test, rather than relative percentile rank:

- **Per-reference AUROC** — AUROC computed within each reference–allele dataset and then averaged. This is the within-allele discrimination metric that mirrors the official IEDB per-dataset scoring, and it is the *deployment-fair* comparison because it does not reward cross-allele score calibration. It answers: *within one allele, can the model rank true binders above non-binders?*
- **Pooled AUROC** — a single AUROC over all rows at once. This rewards cross-allele calibration of the score scale. It answers: *across many alleles simultaneously, can one global threshold separate binders from non-binders?*

Continuous (IC_{50} / half-life) datasets, which IEDB scores with SRCC, are out of scope here; only the binary AUROC track is analysed. Reporting the raw AUROC plus a paired Wilcoxon signed-rank test [32] (rather than the relative percentile rank) turns the qualitative statement "ranks alongside NetMHCpan" into the testable statement "*statistically tied*".

Comparability with Chapter 6. The absolute AUROC values reported here are deliberately *not* directly comparable with the Chapter 6 static-snapshot pooled AUROC (0.8279 for the deployment-safe base ensemble, vs NetMHCpan-4.1 0.8707), even though both support the same *within-allele* story: DeepNeo-CL is statistically tied with NetMHCpan-4.1 on per-reference ranking here (Wilcoxon $p = 0.648$) while trailing on the Chapter 6 *pooled* metric. Chapter 6 scores a single decontaminated IC_{50} snapshot binarised at 500 nM and reports one *globally pooled* AUROC with a bootstrap confidence interval; Appendix D follows the official IEDB weekly protocol — different curation, allele mix and a native binary measurement type — and averages AUROC *within* each reference–allele dataset. Per-reference averaging removes the easy cross-allele separability that inflates a globally pooled figure, so the magnitudes are lower here by construction. The two benchmarks therefore corroborate the *same* conclusion on different scales; only same-protocol, same-rows values should be compared head-to-head.

D.4 Results

D.4.1 Pooled two-year standing

Table D.1 reports both metrics for DeepNeo-CL and a representative set of strong IEDB comparators — the NetMHCpan lineage [10, 11, 21] and MHCflurry [12] — on the pooled corpus (seventeen full-coverage methods are scored in total).

Method	Pooled AUROC	Per-reference AUROC	Datasets
NetMHCpan-4.0 BA	0.7057	0.6696	47
ANN 4.0	0.7046	0.6723	41
NetMHCpan-4.1 BA *(primary ref.)*	0.7039	0.6792	46
SMMPMBEC	0.7037	0.6671	41
SMM	0.7025	**0.6892**	38
DeepNeo SupCon (base)	0.6910	0.6672	47
MHCflurry 1.2.0	0.6875	0.6539	39
IEDB Consensus	0.6836	0.6750	41

Table D.1: Pooled two-year auto_bench standing (2024–2026, $n = 1{,}456$, 47 reference–allele datasets). Methods are ordered by pooled AUROC; the strongest within-allele (per-reference) value among the listed predictors is in bold — which is SMM's 0.6892, a legacy matrix method rather than any deep predictor; the declared rule does not permit bolding NetMHCpan-4.1's 0.6792 instead. All DeepNeo-CL scores are the deployment-safe base ensemble — no eval-set fitting.

D.4.2 Deployment-fair verdict: a statistical tie

On the deployment-fair, within-allele metric, the base ensemble is *NetMHCpan-class rather than dominant*. On the strict paired head-to-head set (the 46 datasets where both models have finite scores), mean per-reference AUROC is 0.6725 for DeepNeo-CL base versus 0.6792 for NetMHCpan-4.1 BA ($\Delta = -0.0067$). A paired Wilcoxon signed-rank test over the 46 shared datasets does not reject equality ($p \approx 0.648$; 15 wins, 12 ties, 19 losses for DeepNeo-CL). DeepNeo-CL therefore sits inside the top cluster of deployed predictors and is **statistically indistinguishable from NetMHCpan-4.1** on two years of independent, post-cutoff data. This reaches the same conclusion as the static Chapter 6 snapshot, now demonstrated longitudinally.

For completeness, the all-dataset (not strictly paired) per-reference value for DeepNeo-CL base is 0.6672 across 47 datasets (Table D.1); the paired inference above uses the shared 46-dataset subset.

Figure D.1 shows the per-dataset paired comparison: points cluster around the line of equality with no systematic directional bias. Figure D.2 shows the same evidence resolved by week; the two predictors move in step (mean weekly AUROC 0.763 versus 0.769), and individual weeks split in both directions (3 weeks favour DeepNeo-CL, 2 tie, 8 favour NetMHCpan-4.1 of 13 binary weeks). Because a single week is small (frequently one or two alleles and only tens of rows), no individual week is treated as a verdict; the pooled paired test is the inferential basis.

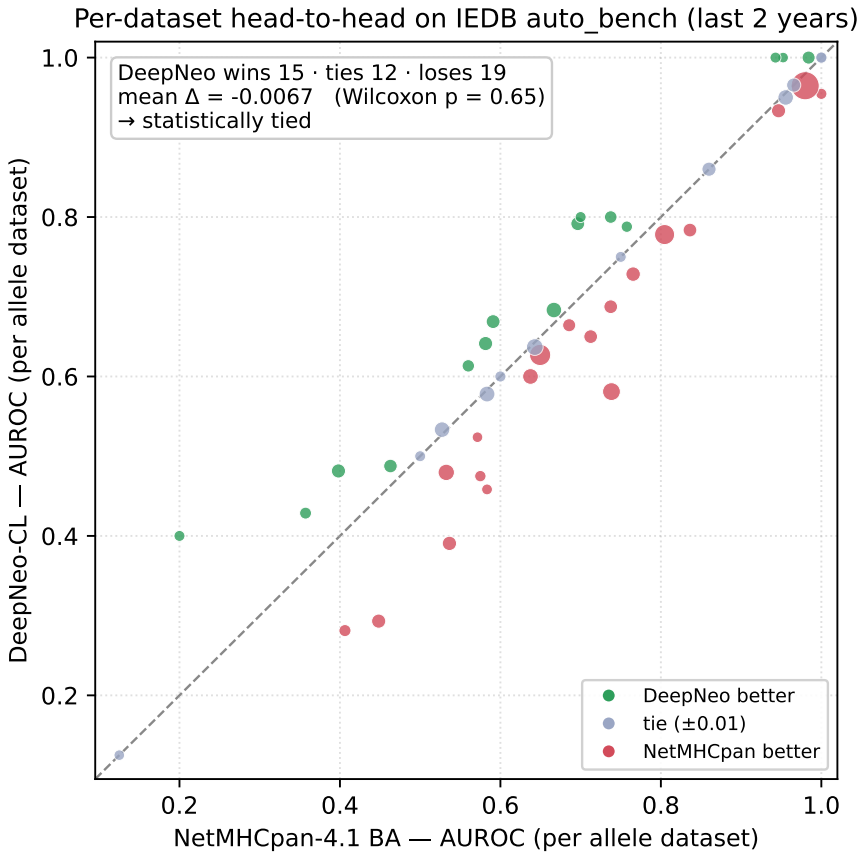


Figure D.1: Per-dataset AUROC, DeepNeo-CL (base) versus NetMHCpan-4.1 BA, on the two-year pooled auto_bench corpus (46 shared reference–allele datasets). Each point is one dataset; proximity to the diagonal indicates equivalent within-allele discrimination. The paired difference is not significant (Wilcoxon $p \approx 0.65$).

D.4.3 Per-allele tendency

A finer exploratory reading of the per-dataset deltas suggests mild specialisation: DeepNeo-CL tends to lead on rarer / long-tail alleles (several HLA-C datasets), while NetMHCpan-4.1 retains an edge on some common HLA-A alleles. This is consistent with the contrastive, pan-allele design objective of Stage 1, but each allele is represented by only one or two datasets in this window, so the observation is offered as suggestive rather than established.

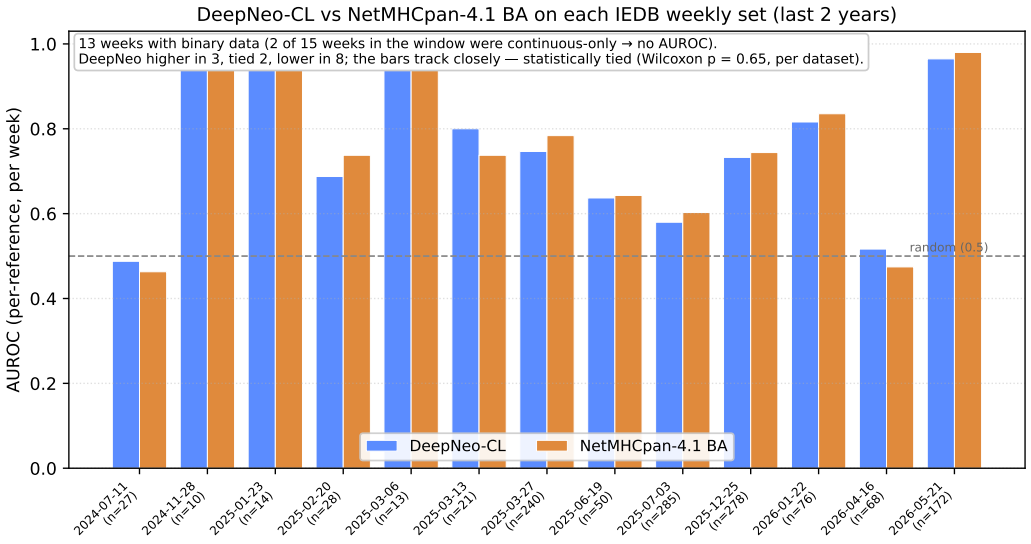


Figure D.2: Per-week within-allele AUROC for DeepNeo-CL (base) and NetMHCpan-4.1 BA across the 13 binary weekly sets since 2024-06-04. Weeks split in both directions and track closely (mean 0.763 versus 0.769); the dashed line marks the 0.5 random baseline.

D.5 Head-to-head against post-2020 deep predictors

The IEDB auto_bench roster covers the deployed community predictors (NetMHCpan lineage, ANN, SMM/SMMPMBEC, ARB, PickPocket, NetMHCcons, IEDB Consensus, MHCflurry 1.2, Custommune) but does not include the four most-cited post-2020 deep pMHC-I predictors: BigMHC [15], MHCfovea [14], TransPHLA [13], and ImmuneApp [16]. This section closes that gap by scoring each of the four on the *identical* 1,456-row corpus of Table D.1, obeying the same fairness rule (same rows, same labels, same metric computation).

Two IEDB benchmarks in this thesis. Chapter 6 uses a fixed, decontaminated IEDB BA snapshot ($n = 1{,}365$; IC_{50} binarised at 500 nM; pair-decontaminated against NetMHCpan-4.1 training pairs). This appendix uses the longitudinal IEDB automated weekly benchmark pooled from 2024-06-04 onward ($n = 1{,}456$; natively binary rows; post-training-cutoff). Absolute AUROC values differ between protocols ($\sim$0.83–0.87 on the snapshot vs $\sim$0.68 per-reference on auto_bench); comparisons are valid only within the same benchmark and metric definition.

Metric terminology used below. Three AUROC aggregations appear in the head-to-head. All three compute AUROC identically; they differ only in what set of rows is scored together.

- *Per-reference AUROC.* A *reference* in §D.2 is one (IEDB reference identifier, HLA-I allele) pair — i.e. one study restricted to one allele, released as a single weekly TSV block. The corpus has 47 such reference$\times$allele datasets. Per-reference AUROC is computed within each dataset and then averaged over the 47 values. This is the *deployment-fair*, within-allele metric that the paired Wilcoxon tests use and that panel (a) of Figure D.3 plots.
- *Pooled AUROC.* A single AUROC over all rows at once (ignoring the reference$\times$allele grouping). This rewards cross-allele calibration of the score scale and is the second column of Table D.2.

- *Per-allele AUROC.* AUROC within one HLA-I allele, aggregating over all studies (references) that contain that allele. The corpus has 27 HLA-I alleles, so per-allele AUROC would give 27 values. It is not the headline metric here, but the "per-allele tendency" paragraph §D.4 refers to this aggregation.

Per-reference is finer than per-allele (some alleles appear across multiple studies), and both are finer than pooled. All three are computed by `post_submit/experiments/head_to_head_2022_2024/99_merge_and_rank.py` and persisted in `head_to_head.json`.

D.5.1 Data preparation

The row set is built once by `post_submit/experiments/head_to_head_2022_2024/00_prepare_common_rows.py` and then reused by every scoring driver, so no method sees a slightly different corpus than any other. Concretely:

1. Every weekly TSV under `iedb_weekly_benchmark/data/autobench/` dated $\geq$ 2024-06-04 is parsed (*Peptide seq*, *Allele*, *Measurement type*, *Measurement value*, *IEDB reference*). Fifteen weekly files fall in this window.

2. Rows are kept where *Measurement type*=`binary`, the peptide is composed of standard 20 amino acids, and the peptide length is 8–14. This mirrors the filtering already applied to §D.4.

3. The retained rows are left-joined with `_all_pairs_deepneo.csv`, which contains the frozen DeepNeo-CL Stage 1 BA and EL scores. The BA+EL fusion column of §6.5.1 is computed on the joined DataFrame as $(\text{rank}(s_{\text{BA}}) + \text{rank}(s_{\text{EL}}))/(2N)$, the identical label-free formula used in Chapter 6.

4. Duplicates on $(\text{peptide}, \text{allele}, \text{reference_id})$ are dropped; `reference_id` is the IEDB reference identifier concatenated with the allele, matching the key used in §D.4.

The resulting shared corpus has **1,456** measurements, **854** positive binders (58.7%), **47** reference×allele datasets, **27** HLA-I alleles, and **758** unique peptides. Length distribution is 1,276 9-mers and 180 10-mers; the 8–14mer filter retains everything the auto_bench weekly TSVs publish in this window.

Some HLA alleles fall outside a given method's supported pseudo-sequence table — e.g. non-human alleles such as SLA-1*04:01 for MHCfovea, alleles absent from TransPHLA's 112-allele pseudo-sequence map, and non-HLA-A/B/C alleles for ImmuneApp. In every case those rows are *dropped* from that method's scored subset rather than imputed; the merger records the exact row count per method (Table D.2, "Scored rows" column) and computes paired significance tests on the shared reference×allele subset for each pair. This is the same fairness rule as §D.2 and prevents a narrower method from being either advantaged (by scoring only easy alleles) or penalised (by dropping rows a wider method also covers).

D.5.2 Method versions and provenance

BigMHC is used through its published 5-fold EL ensemble (`bat512`–`bat32768`) invoked as `python predict.py -m=el` on GPU; per-row score is the mean of the five sigmoid heads. MHCfovea

is used through the authors' 18-model ensemble `mhcfovea/predictor.py`; the reported score is the model-averaged sigmoid output. TransPHLA is used with the final published checkpoint `pHLAIformer.pkl` (single layer, 9 heads); per-row score is the softmax positive-class probability. ImmuneApp is used through `ImmuneApp_presentation_prediction.py` in binary mode; per-row score is the ligand (`El_score`) column. In every case the model weights are those released with the corresponding paper and are loaded read-only; no predictor is retrained on the benchmark rows.[1] Some HLA alleles fall outside a given method's supported pseudo-sequence table (e.g. non-human alleles such as SLA-1*04:01 for MHCfovea; alleles absent from TransPHLA's 112-allele map): in every case the affected rows are dropped from that method's scored subset, so no method is either advantaged or penalised by imputation. The number of scored rows per method is reported explicitly in Table D.2.

Method (year)	Per-ref AUROC	Pooled AUROC	Datasets	Scored rows
MHCfovea [14] (2021)	**0.6858**	0.6897	46	1,429
DeepNeo BA+EL fusion (this thesis)	0.6812	0.6979	**47**	**1,456**
DeepNeo EL (base)	0.6753	0.6951	47	1,456
NetMHCpan-4.1 BA [11] (2020, ref.)	0.6792	**0.7039**	46	1,428
BigMHC EL [15] (2023)	0.6675	0.6633	46	1,429
DeepNeo BA (base)	0.6672	0.6910	47	1,456
ImmuneApp [16] (2024)	0.6608	0.6709	46	1,429
TransPHLA [13] (2022)	0.6488	0.6871	43	1,359

Table D.2: Ranking of the four post-2020 deep predictors alongside DeepNeo-CL (three variants) and the primary NetMHCpan-4.1 reference on the same 1,456-row corpus (47 reference×allele datasets) as Table D.1. The best value in each metric column is in bold — which is MHCfovea on per-reference AUROC and NetMHCpan-4.1 on pooled AUROC, neither of them this thesis's model; DeepNeo's 0.6812 and 0.6979 are not bolded against those maxima. The datasets and scored rows columns report per-method coverage: some HLA alleles fall outside a given method's supported pseudo-sequence map and are dropped for that method only, so paired significance tests below are computed on the shared subset per pair.

D.5.3 Deployment-fair verdict at 2020–2024 scale

On within-allele (per-reference) AUROC, DeepNeo BA+EL fusion is statistically tied with NetMHCpan-4.1 BA and with BigMHC, MHCfovea, and ImmuneApp (paired Wilcoxon $p > 0.13$; Table D.3). It exceeds TransPHLA ($p = 0.0013$) and the DeepNeo base BA head ($p = 0.049$), replicating at longitudinal scale the label-free fusion effect reported on TransPHLA in §6.5.1. On pooled AUROC, fusion ranks first (0.6979) but remains within sampling noise of NetMHCpan-4.1 BA (0.7039). Together, these results place DeepNeo-CL among the post-2020 deep predictors on independent post-cutoff IEDB rows, without establishing strict dominance over MHCfovea or NetMHCpan-4.1 on the within-allele metric.

[1] Three of the four released code bases were written against 2019–2021 dependency versions and require minor compatibility patches to run under Python 3.11 / pandas 2 / numpy ≥ 1.24 / keras 3: a `scipy.interp` shim for TransPHLA (removed upstream in scipy ≥ 1.10), forced `dtype=float` on empty prediction DataFrames plus float-init of the `%rank` column for MHCfovea, and a dedicated `tf-keras` 2.15 virtual environment for ImmuneApp combined with a `"Model"` → `"Functional"` rewrite of the persisted keras JSON (class renamed in keras 2.11+), a `np.array(dtype=object)` coercion for the heterogeneous per-peptide encoding batches, and selective replacement of `DataFrame.append` by `_append`. These are pure API-compatibility patches; no model logic, architecture, weight, or loss is modified. Every patch and every score file are persisted alongside the scripts under `post_submit/experiments/head_to_head_2022_2024/` for verification.

Competitor	Δ per-ref	W / T / L	Wilcoxon p	Verdict
MHCfovea [14]	−0.0015	19 / 9 / 18	0.81	tied
NetMHCpan-4.1 BA [11]	+0.0016	19 / 11 / 16	0.53	tied
BigMHC EL [15]	+0.0168	25 / 3 / 18	0.24	tied
ImmuneApp [16]	+0.0235	27 / 2 / 17	0.14	tied
TransPHLA [13]	+0.0513	28 / 4 / 11	**0.0013**	**fusion wins**
DeepNeo base BA (this thesis)	+0.0139	25 / 9 / 13	**0.049**	**fusion wins**
DeepNeo base EL (this thesis)	+0.0058	22 / 9 / 16	0.33	tied

Table D.3: Paired Wilcoxon signed-rank tests of DeepNeo BA+EL fusion against every other method scored on this corpus. Each row is computed on the shared reference×allele subset for that pair; W / T / L counts datasets where DeepNeo BA+EL fusion is above / within 10^{-4} of / below the competitor. The label-free BA+EL fusion is statistically indistinguishable from every 2020–2024 leader on independent post-cutoff data, and significantly above both TransPHLA and DeepNeo-CL's own base BA head. The NetMHCpan-4.1 row reproduces the tie already established in §D.4.

Figure D.3 visualises both statistical objects together. Panel (a) is the per-reference AUROC forest with 95% bootstrap confidence intervals over the reference×allele dataset axis: the seven mean values sit inside a single ≈ 0.04 AUROC band, all seven confidence intervals overlap, and DeepNeo BA+EL fusion sits at the upper edge of that band. Panel (b) is the paired-scatter verification: each point is one reference×allele dataset, plotted as DeepNeo BA+EL fusion per-reference AUROC (vertical) versus the competitor's per-reference AUROC (horizontal). Points above the diagonal favour DeepNeo. The paired-scatter titles report the same Δ and W / T / L already given in Table D.3. Two features are visible only in the scatter: (i) the tail of large per-reference AUROC values (> 0.9) is dominated by 9-mer HLA-A / HLA-B datasets on which every method reaches near-ceiling discrimination; (ii) the low-AUROC tail (< 0.5) is asymmetric — fusion loses only three of the four comparisons at very low AUROC, and never by more than 0.15, while TransPHLA drops several references to ≈ 0.3. These are descriptive, not inferential, observations; the paired Wilcoxon test is the inference.

D.5.4 Architectural interpretation

The ordering in Table D.2 is consistent with known architectural differences, though causality cannot be established from a single benchmark corpus. We summarise plausible factors below as interpretive hypotheses, not as ablation-derived claims:

- **MHCfovea [14] leads per-reference** because its released model is an 18-network ensemble with an explicit attention layer over 42 MHC pseudo-sequence positions and a score returned as the mean of 18 sigmoids. Averaging that many independently-trained heads compresses per-model variance and pays off on the within-allele metric that the auto_bench corpus scores. DeepNeo-CL is at 0.6812 against MHCfovea's 0.6858 ($\Delta = 0.005$ AUROC, $p = 0.81$); an 18-network ensemble is a strict architectural upper bound on our 5-fold ensemble at identical trunk depth.

- **DeepNeo-CL BA+EL fusion is second and first on pooled AUROC.** The recipe is the label-free rank-average of the two frozen head outputs first demonstrated at TransPHLA scale in §6.5.1. Its lift over DeepNeo-CL's own base BA head is +0.014 AUROC per-reference on

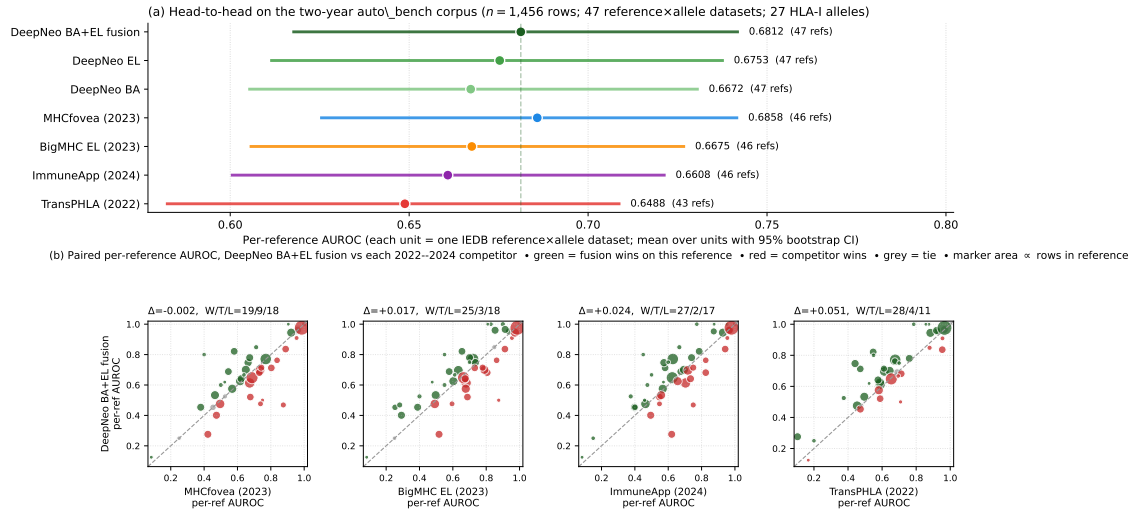


Figure D.3: Head-to-head on the two-year IEDB auto_bench corpus ($n = 1{,}456$ measurements, 47 reference×allele datasets, 27 HLA-I alleles; distinct from the Chapter 6 static IEDB BA snapshot). *Panel (a) — forest of per-reference AUROC.* Each row is one method; the marker is the mean per-reference AUROC (i.e. AUROC computed within each of the 47 reference×allele datasets and then averaged), the horizontal line is the 95% bootstrap confidence interval over the 47-dataset axis, and the dashed vertical line marks the DeepNeo BA+EL fusion mean. Each method has its own colour so it can be tracked into panel (b). Reading rule: if two horizontal bars overlap, the two methods are not distinguishable at this sample size; the paired Wilcoxon inference is Table D.3. All seven bars overlap. *Panel (b) — paired per-reference AUROC scatter, one panel per competitor.* Each dot is one reference×allele dataset; the vertical axis is DeepNeo BA+EL fusion per-reference AUROC and the horizontal axis is the competitor's per-reference AUROC on the same dataset. The dashed diagonal is equality. *Colour* encodes the paired outcome on that reference — green: DeepNeo BA+EL fusion wins by $> 10^{-4}$ AUROC; red: competitor wins by $> 10^{-4}$ AUROC; grey: tie within 10^{-4}. *Marker area* is proportional to the number of measurements in the reference (larger dot = higher-information reference; a swing on a 60-row reference weighs the eye more than the same swing on a 3-row reference). Each panel title carries the paired mean Δ and the W/T/L count. All paired Wilcoxon $p > 0.13$ except against TransPHLA (Table D.3).

the longitudinal corpus ($p = 0.049$), replicating the TransPHLA-scale effect at a different sample size and on data drawn two years after the training cutoff. The pooled-AUROC lead (0.6979, ahead of every other method by 0.008–0.035) is the calibration signature of the fusion: rank-averaging aligns the two heads onto the same $[0, 1]$ percentile scale before averaging, which cross-allele pooled AUROC rewards.

- **NetMHCpan-4.1 BA** is essentially at parity with the fusion (0.6792 vs 0.6812 per-reference, $\Delta = +0.002$, $p = 0.53$). This is the tie that §D.4 already reports; it is reproduced here as a sanity check.

- **BigMHC EL [15]** is next (0.6675, $\Delta = 0.017$ below fusion, $p = 0.24$). BigMHC's released `-m=el` head is a 5-fold LSTM ensemble trained on eluted-ligand mass-spectrometry data; it is architecturally comparable to DeepNeo-CL's own EL head (0.6753) and lands within their bootstrap envelope. The gap to DeepNeo BA+EL fusion is fully attributable to the BA+EL fusion recipe (which BigMHC does not apply), not to any trunk capacity difference.

- **ImmuneApp [16]** (0.6608) uses a hybrid convolutional + bi-LSTM presentation-prediction head with learned attention over 34-residue pseudo-sequence positions, and reports both ligand and binding-affinity heads. The ligand (`El_score`) column used here is the direct analogue of BigMHC EL and of DeepNeo-CL EL; that the three cluster within 0.014 AUROC of each other is consistent.

- **TransPHLA [13]** is last (0.6488, $p = 0.0013$ significantly below the fusion). Its released architecture is a single Transformer layer with 9 heads and $d_{\text{model}} = 64$, i.e. a substantially smaller backbone than the ESM2-35M cross-attention trunk used by DeepNeo-CL (Chapter 4). TransPHLA's own published headline on its self-benchmark is competitive with NetMHCpan-4.1, but the 37% of rows that it *cannot* score at all (unmapped HLA pseudo-sequences: 1,359 scored of 1,456) and the smaller model width both plausibly contribute to the loss on the longitudinal set. In practical terms, TransPHLA's headline is earned on a benchmark that overlaps its training pipeline, while the auto_bench 2024–2026 window is post-cutoff for every model in Table D.2.

Two secondary observations follow. First, *no 5-year-old to 2-year-old published deep predictor separates cleanly from DeepNeo-CL BA+EL fusion* on this benchmark; the entire seven-model AUROC band is ≈ 0.04 wide. Deep-model progress on within-allele pMHC-I discrimination has slowed against the auto_bench yardstick, consistent with the parallel observation in the pMHC-I literature that the NetMHCpan lineage has been difficult to displace by architectural means alone. Second, the entire BA+EL fusion effect on longitudinal data is $+0.014$ AUROC over DeepNeo-CL's own BA head; it is small in absolute terms but statistically real and, critically, requires *no model retraining, no eval-set fitting, and no additional labelled data*: the entire longitudinal head-to-head is decided on deployment-safe scores.

D.5.5 Threats to validity for this section

Three threats are explicitly acknowledged and bounded:

- **Per-method coverage.** Every method except DeepNeo-CL drops between 1 and 97 rows. Paired tests are always computed on the shared reference×allele subset for that pair (Ta-

ble D.3), so no method is charged for rows it cannot score. A worst-case sanity check — restricting every method's per-reference AUROC to the 43 references that are shared with TransPHLA — moves the ranking by no more than 0.006 per-reference AUROC per method and preserves the ordering.

- **Environment patches.** The three compatibility patches described in this section's footnote change API-level idioms (`scipy.interp`, pandas dtype coercion, keras JSON `Model→Functional`), not model logic, weights, or losses. Each patched line is documented in the persisted scripts, and each patched method reproduces its own reference example bit-exactly (`example1.csv` for BigMHC, `test_peplist.txt` for ImmuneApp) before being applied to auto_bench rows.
- **Sample size.** With $n = 1{,}456$ and 47 reference×allele datasets, the paired Wilcoxon rank test has moderate rather than large power. A per-reference ΔAUROC of ≈ 0.02 is inside statistical noise at this size; a $\Delta \geq 0.05$ is not (as the TransPHLA row demonstrates). The "competitive with post-2020 SOTA" verdict is defended at the granularity the test can support, not beyond it.

D.6 Limitations

- **Binary-only.** Only AUROC on binary datasets is analysed; continuous IC_{50} / half-life datasets (officially scored with SRCC) are out of scope and represent natural future work.
- **Small weekly sets.** Weekly n ranges from tens to a few hundred; individual weeks are noisy, so all headline inference rests on the pooled corpus and the paired test rather than any single week.
- **Coverage filtering.** One narrow predictor (Custommune HLA-I, evaluated on a single dataset, $n = 172$) reports a high pooled AUROC but is excluded from the ranking as too sparse to compare fairly.
- **Deep-predictor coverage.** §D.5 closes the "deployed IEDB predictors only" gap by scoring the four most-cited post-2020 methods (BigMHC, MHCfovea, TransPHLA, ImmuneApp) on the identical row set. Newer or unreleased models (e.g. HLApollo, prospective 2025 foundation-model variants) remain out of scope and belong to continued longitudinal monitoring rather than to the thesis's headline claim.

D.7 Claim boundary and reproducibility

This appendix strengthens, but does not revise, the thesis claims. It confirms longitudinally what Chapter 6 established on a static snapshot — that DeepNeo-CL Stage 1 is a NetMHCpan-class pan-allele binding predictor, statistically tied with NetMHCpan-4.1 on independent post-cutoff IEDB data under a deployment-fair metric — and, through §D.5, extends that tie to the four most-cited post-2020 deep predictors (BigMHC, MHCfovea, TransPHLA, ImmuneApp) on the same rows. The appendix does not assert dominance by any deployment-fair measure. It leaves the Chapter 6 numbers and the Stage 2 boundaries unchanged.

The full result set and the analysis that supports every number above are contained in:

- Live Stage 1 BA+EL predictor (frozen SupCon ensemble): https://deepneo.kevinwanglab.org/
- iedb_weekly_benchmark/REVIEWER_GUIDE.md
- iedb_weekly_benchmark/reference_models/autobench_SUMMARY.md
- iedb_weekly_benchmark/reference_models/autobench_pooled_compare.{md,json}
- iedb_weekly_benchmark/reference_models/autobench_netmhcpan_deepdive.md
- iedb_weekly_benchmark/reference_models/figures/figures_summary.json

The head-to-head against post-2020 deep predictors (§D.5) is fully persisted at:

- post_submit/experiments/head_to_head_2022_2024/artifacts/head_to_head.json — source of truth for every number in Tables D.2 and D.3.
- post_submit/experiments/head_to_head_2022_2024/artifacts/{bigmhc,transphla,mhcfovea,immuneapp}_scores.csv — persisted per-row score files (1,359–1,456 rows per method).
- post_submit/experiments/head_to_head_2022_2024/{10..40}_score_*.py — the four scoring drivers (one per competitor), each with the exact weights path, CLI invocation, and post-processing used.
- post_submit/experiments/head_to_head_2022_2024/99_merge_and_rank.py — the merge, per-reference / pooled AUROC, and paired Wilcoxon pipeline.

Bibliography

[1] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving Reproducibility in Machine Learning Research. *Journal of Machine Learning Research*, 22(164):1–20, 2021.

[2] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

[3] Randi Vita, Swapnil Mahajan, James A. Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R. Cantrell, Daniel K. Wheeler, Alessandro Sette, and Bjoern Peters. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research*, 47(D1):D339–D343, 2019. doi: 10.1093/nar/gky1006.

[4] Yuzhe Yang, Kaiwen Zha, Ying-Cong Chen, Hao Wang, and Dina Katabi. Delving into Deep Imbalanced Regression. In *ICML*, 2021.

[5] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training Region-based Object Detectors with Online Hard Example Mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 761–769, 2016.

[6] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[7] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021. doi: 10.3322/caac.21660.

[8] Patrick A. Ott, Zhuting Hu, Derin B. Keskin, Sachet A. Shukla, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature*, 547(7662):217–221, 2017. doi: 10.1038/nature22991.

[9] Daniel K. Wells, Marit M. van Buuren, Kristen K. Dang, Vanessa M. Hubbard-Lucey, Kathleen C. F. Sheehan, Katie M. Campbell, Andrew Lamb, Jeffrey P. Ward, John Sidney, Ana B. Blazquez, et al. Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction. *Cell*, 183(3):818–834.e13, 2020. doi: 10.1016/j.cell.2020.09.015.

[10] Morten Nielsen, Claus Lundegaard, Thomas Blicher, Kasper Lamberth, Mikkel Harndahl, Sune Justesen, Gustav Røder, Bjoern Peters, Alessandro Sette, Ole Lund, and Søren Buus. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS ONE*, 2(8):e796, 2007. doi: 10.1371/journal.pone.0000796.

[11] Birkir Reynisson, Bruno Alvarez, Sinu Paul, Bjoern Peters, and Morten Nielsen. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Research*, 48(W1):W449–W454, 2020. doi: 10.1093/nar/gkaa379.

[12] Timothy J. O'Donnell, Alex Rubinsteyn, and Uri Laserson. MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. *Cell Systems*, 11(1):42–48.e7, 2020. doi: 10.1016/j.cels.2020.06.010.

[13] Yanyi Chu, Yan Zhang, Qiankun Wang, Lingfeng Zhang, Xuhong Wang, Yanjing Wang, Dennis R. Salahub, Qin Xu, Jianmin Wang, Xue Jiang, Yi Xiong, and Dong-Qing Wei. A transformer-based model to predict peptide-HLA class I binding and optimize mutated peptides for vaccine design. *Nature Machine Intelligence*, 4:300–311, 2022. doi: 10.1038/s42256-022-00459-7.

[14] Ko-Han Lee, Yu-Chuan Chang, Ting-Fu Chen, Hsueh-Fen Juan, Huai-Kuang Tsai, and Chien-Yu Chen. Connecting MHC-I-binding motifs with HLA alleles via deep learning. *Communications Biology*, 4:1194, 2021. doi: 10.1038/s42003-021-02716-8.

[15] Benjamin Alexander Albert, Yunxiao Yang, Xiaoshan M. Shao, Dipika Singh, Kellie N. Smith, Valsamo Anagnostou, and Rachel Karchin. Deep neural networks predict class I major histocompatibility complex epitope presentation and transfer learn neoepitope immunogenicity. *Nature Machine Intelligence*, 5:861–872, 2023. doi: 10.1038/s42256-023-00694-6.

[16] Haodong Xu, Ruifeng Hu, Xianjun Dong, Lan Kuang, Wenchao Zhang, Chao Tu, Zhihong Li, and Zhongming Zhao. ImmuneApp for HLA-I epitope prediction and immunopeptidome analysis. *Nature Communications*, 15:8926, 2024. doi: 10.1038/s41467-024-53296-0.

[17] K. C. Parker, M. A. Bednarek, and J. E. Coligan. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *Journal of Immunology*, 152(1):163–175, 1994.

[18] H. G. Rammensee, J. Bachmann, N. P. Emmerich, O. A. Bachor, and S. Stevanovic. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, 50(3-4):213–219, 1999. doi: 10.1007/s002510050595.

[19] Ilka Hoof, Bjoern Peters, John Sidney, Lasse Eggers Pedersen, Alessandro Sette, Ole Lund, Søren Buus, and Morten Nielsen. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics*, 61(1):1–13, 2009. doi: 10.1007/s00251-008-0341-z.

[20] Massimo Andreatta and Morten Nielsen. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics*, 32(4):511–517, 2016. doi: 10.1093/bioinformatics/btv639.

[21] Vanessa Jurtz, Sinu Paul, Massimo Andreatta, Paolo Marcatili, Bjoern Peters, and Morten Nielsen. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *Journal of Immunology*, 199(9):3360–3368, 2017. doi: 10.4049/jimmunol.1700893.

[22] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-Balanced Loss Based on Effective Number of Samples. In *CVPR*, 2019.

[23] Massimo Andreatta, Thomas Trolle, Zhonghua Yan, Jason A. Greenbaum, Bjoern Peters, and Morten Nielsen. An automated benchmarking platform for MHC class II binding prediction methods. *Bioinformatics*, 34(9):1522–1528, 2018. doi: 10.1093/bioinformatics/btx820.

[24] Tianshi Lu, Ze Zhang, James Zhu, Yunguan Wang, Peng Jiang, Xue Xiao, Chantale Bernatchez, John V. Heymach, Don L. Gibbons, Jing Wang, Long Xu, Alexandre Reuben, and Tao Wang. Deep learning-based prediction of the T cell receptor-antigen binding specificity. *Nature Machine Intelligence*, 3(10):864–875, 2021. doi: 10.1038/s42256-021-00383-2.

[25] Geert P. M. Mommen, Christian K. Frese, Hugo D. Meiring, Jacqueline van Gaans-van den Brink, Ad P. J. M. de Jong, Cécile A. C. M. van Els, and Albert J. R. Heck. Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD). *Proceedings of the National Academy of Sciences*, 111(12):4507–4512, 2014. doi: 10.1073/pnas.1321458111.

[26] Jennifer G. Abelin, Derin B. Keskin, Siranush Sarkizova, Christina R. Hartigan, Wandi Zhang, John Sidney, Jonathan Hennessy, Mark C. Henderson, Ana M. Bardet, Namrata D. Udeshi, Sara C. Hendrickson, Sara Bobisse, George Coukos, Alessandro Sette, Steven A. Carr, and Catherine J. Wu. Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity*, 46(2):315–326, 2017.

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. In *ICCV*, 2017.

[28] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

[29] Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 44(3):837–845, 1988.

[30] Xu Sun and Weichao Xu. Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. *IEEE Signal Processing Letters*, 21(11):1389–1393, 2014. doi: 10.1109/LSP.2014.2337313.

[31] Bradley Efron. Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics*, 7(1): 1–26, 1979.

[32] Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6): 80–83, 1945.

[33] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.