CAPA: Directional, Structure-Aware HLA Mismatch Representations from Protein Language Models for Competing-Risks Survival Analysis in Haematopoietic Stem Cell Transplantation

Original Paper · *Bioinformatics*

Huanxuan Li[1,*]

[1]Thomas Jefferson High School for Science and Technology, Alexandria, United States
[*]Correspondence: 2027hli@tjhsst.edu

Abstract

Motivation: Donor selection for allogeneic haematopoietic stem cell transplantation (HSCT) relies on categorical HLA match/mismatch counts that treat antigenically distinct alleles as equivalent once they share a mismatch flag. This representation discards the amino-acid-level structural divergence between allele pairs and ignores the competing nature of post-transplant events: graft-versus-host disease (GvHD), relapse, and transplant-related mortality (TRM) preclude one another and must be modelled jointly. A subtler limitation affects even the continuous distance metrics proposed as replacements: a Euclidean mismatch distance is symmetric and therefore *direction-blind*, whereas graft-versus-host alloreactivity is intrinsically directional. This work asks whether a representation learned end-to-end can capture directional mismatch information that no scalar distance can encode.

Methods: We introduce CAPA (Computational Architecture for Predicting Alloimmunity), a deep learning framework that encodes each HLA allele as a 1280-dimensional vector using the frozen ESM-2 650M protein language model. Donor–recipient interaction features are extracted by a bidirectional cross-attention network (2 layers, 8 heads), which is concatenated with clinical covariates and passed to a DeepHit head that jointly estimates the discrete-time cumulative incidence functions for all three competing events over a 730-day horizon. Only the interaction network and survival head (∼27.8M parameters) are trained by default; ESM-2 is frozen but can optionally be partially fine-tuned via a second AdamW parameter group. Post-hoc isotonic calibration is applied per event and time bin to improve CIF reliability. A multi-donor comparison endpoint ranks up to 20 candidates by composite GvHD+TRM risk. Because the full cross-attention model is under-determined at the cohort sizes available here, the end-to-end experiments use a compact 469K-parameter variant that self-attends over per-locus *signed* difference embeddings $\mathbf{e}^{\mathrm{D}} - \mathbf{e}^{\mathrm{R}}$; we report this variant as CAPA throughout the simulation results. We evaluate the framework on the public UCI BMT cohort, a real-world allele-level dataset (IHWG), and three controlled simulations designed to isolate *what* the learned representation adds over scalar mismatch distances; we do not claim a clinically validated predictor.

Results: On the public UCI Bone Marrow Transplant dataset ($n = 187$ paediatric patients, train/val/test 70/15/15 %), we benchmark tabular-feature competing-risks models as reference comparators for CAPA. Because the held-out test set is small ($n = 29$), we treat repeated 5×5 cross-validation as the primary performance estimate: the cause-specific Cox model reaches relapse $C = 0.60 \pm 0.14$ and TRM 0.56 ± 0.06. Single held-out-split numbers are markedly more optimistic (Fine–Gray relapse 0.84, TRM 0.66; Cox 0.75/0.65) but carry very wide intervals (e.g. relapse 95 % CI 0.69–1.00 on four test events) and should not be

over-interpreted; a Random Survival Forest (relapse 0.48) and a flat-feature DeepHit MLP with a correctly formulated censored log-likelihood (relapse 0.65, TRM 0.57) complete the reference set. A feature ablation study shows that aggregate HLA mismatch scores alone yield near-chance concordance ($C \approx 0.49$), while clinical features alone match the full-feature model, directly motivating ESM-2 structural allele representations as the form of HLA information most likely to add discriminative signal. A controlled mechanistic simulation provides a capability check: in a cohort where all patients share an identical binary DRB1 mismatch profile, a Cox model restricted to binary mismatch achieves GvHD $C = 0.501$ (chance), whereas a Cox model given the continuous ESM-2 alloreactivity distances achieves $C = 0.695$ ($\Delta C = +0.194$)—confirming that a continuous representation recovers a planted signal invisible to binary indicators, though the signal is generated from the distances by construction and the experiment is therefore a sanity check rather than evidence of biological informativeness. A scale-up simulation in the haploidentical setting ($N = 10{,}000$; all five loci mismatched) extends this result: Cox (ESM-2 distances) achieves $C_{\text{GvHD}} = 0.759$ vs. 0.538 for binary mismatch ($\Delta C = +0.221$). CAPA (locus attention, $d_{\text{proj}} = 64$) achieves $C_{\text{GvHD}} = 0.732$, statistically indistinguishable from Cox (ESM-2 distances) (95 % CIs overlap), demonstrating that alloreactivity self-attention recovers the main-effect signal from raw embeddings without explicit distance features. Crucially, a scalar mismatch distance is symmetric and magnitude-only ($\|\mathbf{e}^{\text{D}} - \mathbf{e}^{\text{R}}\| = \|\mathbf{e}^{\text{R}} - \mathbf{e}^{\text{D}}\|$) and is therefore *blind to the direction* of donor–recipient mismatch, whereas graft-versus-host alloreactivity is intrinsically directional. In a directional GvHD simulation (six seeds), the scalar-distance Cox model collapses to near-chance ($C_{\text{GvHD}} = 0.575 \pm 0.059$) while CAPA—learning the immunodominant direction end-to-end from the *signed* difference embeddings—reaches $C_{\text{GvHD}} = 0.869 \pm 0.023$, a gain of $\Delta C = +0.294$ that recovers 93 % of the gap to a direction-aware oracle (0.892 ± 0.015), with non-overlapping confidence intervals on every seed. This identifies the learned representation's genuine architectural advantage: it captures directional alloreactivity that no scalar distance can encode. Real-world validation on the publicly available IHWG HCT dataset ($n = 1{,}347$, allele-level HLA typing, OS endpoint) confirms that ESM-2 distances are *comparable* to binary mismatch indicators overall ($C = 0.602$ vs 0.615); in the targeted subgroup of single-mismatch non-CML patients ($n = 112$), ESM-2 distances exceed binary features ($\Delta C = +0.024$), directionally consistent with the mechanistic prediction. Registry-scale data with per-allele HLA typing and event-specific endpoints (CIBMTR) is identified as the primary direction for full validation.

Availability and Implementation: Source code, pre-trained weights, and a web interface (including multi-donor comparison) are freely available at https://github.com/sh4wn27/capa under the MIT licence.

Keywords: HLA; protein language model; competing risks; haematopoietic stem cell transplantation; graft-versus-host disease; deep learning; survival analysis; cross-attention

1 Introduction

Allogeneic haematopoietic stem cell transplantation (HSCT) is the only curative treatment for many high-risk haematological malignancies, including acute leukaemia, myelodysplastic syndromes, and severe aplastic anaemia [3, 30]. Global HSCT activity now exceeds 50,000 procedures annually, yet five-year overall survival for unrelated-donor transplants remains below 50 % in many disease categories, underscoring the need for improved outcome prediction at the time of donor selection [18, 30].

The three dominant causes of treatment failure after HSCT—acute graft-versus-host disease (aGvHD), disease relapse, and transplant-related mortality (TRM)—are *competing risks* [14]. aGvHD, driven by donor T-cell recognition of recipient alloantigens, affects 30–70 % of recipients and accounts for up to 15 % of all transplant deaths [13, 41]. Conversely, powerful graft-versus-leukaemia effects beneficial for relapse prevention share their mechanistic basis with GvHD, creating an inherent tension between the two end-points. Because occurrence of any one event precludes or substantially alters the hazard of the others, standard time-to-event models fitted independently for each outcome overestimate cumulative incidence and yield statistically biased risk predictions [14, 20]. Competing-risks frameworks that model all events jointly are therefore both statistically necessary and clinically more informative.

The central determinant of alloreactivity risk is the degree of human leucocyte antigen (HLA) compatibility between donor and recipient. HLA class I molecules (HLA-A, -B, -C) and class II molecules (HLA-DRB1, -DQB1) display intracellular peptides on the cell surface; polymorphisms in their peptide-binding grooves determine which self and foreign epitopes are presented and whether donor T cells recognise recipient cells as foreign [38]. Current clinical practice evaluates compatibility by counting the number of mismatched loci at high-resolution allele-level typing ("10/10" or "8/8" matching), with each additional mismatch conferring incrementally worse outcomes at a population level [23, 31].

This locus-count approach has two structural limitations. First, it treats all mismatches as immunologically equivalent regardless of the physicochemical distance between allele products: two alleles that differ by a single conservative substitution at a solvent-exposed position are penalised identically to two alleles that differ at multiple immunodominant residues in the peptide-binding groove [7, 15, 32]. Attempts to refine categorical scoring using amino-acid-level electrostatic mismatch scores [37] or structural "eplet" classifications [11] have shown modest clinical utility but remain hand-engineered and locus-specific, limiting generalisability. Second, aggregate mismatch counts lose information about *which* loci are mismatched and the direction of mismatch (donor-versus-host versus host-versus-donor), both of which influence event-specific risk [15, 32].

Protein language models (PLMs) offer a principled, data-driven alternative to hand-crafted mismatch scores [12, 24, 34]. Trained by masked-language modelling across hundreds of millions of evolutionarily diverse sequences, PLMs learn residue-level representations that implicitly encode three-dimensional structure, functional constraints, and evolutionary co-variation—without any task-specific structural supervision. The ESM-2 model family has demonstrated that these representations are predictive of mutational fitness effects [29], protein–protein interaction sites [34], and, via ESMFold, atomic-resolution structure [24]. Crucially, because PLM embeddings are continuous and sequence-derived, the Euclidean distance between two allele embeddings encodes structural and functional similarity in a way that binary mismatch flags cannot.

Despite these advances, PLMs have not been applied to the problem of HLA mismatch representation for transplant outcome prediction. Existing machine-learning approaches to HSCT outcome modelling treat HLA information as a categorical covariate—typically a per-locus mismatch binary or an aggregate mismatch count—and focus modelling capacity on clinical

covariates [10, 35]. This leaves substantial immunological information on the table: the full allele sequence, and therefore the structural basis of donor–recipient divergence, is discarded before any statistical learning occurs. Furthermore, existing HSCT prediction models—whether logistic regression, random forest, or early deep learning [10, 35]—treat GvHD, relapse, and TRM as separate binary or time-to-event targets, ignoring the statistical dependency imposed by their competing-risks relationship. Analysing competing events with independent models yields biased cumulative incidence estimates and misrepresents the absolute probability that a patient will experience each outcome by a given time [14, 33]. Jointly modelling all three end-points as a competing-risks process is therefore both statistically necessary and clinically more informative: it directly estimates the probability a patient will experience each specific event before the others, which is the quantity relevant to donor selection.

Here we introduce **CAPA** (Computational Architecture for Predicting Alloimmunity), a deep learning framework that addresses both limitations simultaneously. CAPA (i) replaces categorical mismatch flags with 1280-dimensional ESM-2 embeddings of HLA allele protein sequences; (ii) learns donor–recipient alloreactivity end-to-end via attention over the *signed* difference embeddings, capturing the *direction* of HLA mismatch—which antigens the recipient carries that the donor lacks—a property that the symmetric, magnitude-only scalar distances used in current practice are structurally incapable of representing; and (iii) estimates all three competing risks jointly through a DeepHit survival head [22] that produces calibrated cumulative incidence functions over a 730-day post-transplant horizon. We describe the CAPA architecture, implement and release the full pipeline, and benchmark tabular-feature competing-risks baselines (cause-specific Cox, Fine–Gray, DeepHit MLP) on the publicly available UCI Bone Marrow Transplant (BMT) paediatric dataset [36] to establish reference performance achievable without allele-level HLA data. The UCI BMT dataset records aggregate mismatch scores rather than specific allele identities, precluding end-to-end evaluation of the ESM-2 embedding step; validation on a registry dataset with high-resolution HLA typing constitutes the primary direction for future work. We position this paper accordingly as a *representation-level proof of concept*: our central, fully supported claim is that a learned representation captures directional alloreactivity that scalar mismatch distances provably cannot (Section 3.11); a clinically validated outcome predictor would require registry-scale data with allele-level typing and event-specific endpoints, which we identify but do not claim to deliver.

2 Methods

Figure 1 provides an overview of the CAPA architecture. The framework comprises four main components: (i) an ESM-2 650M protein language model that encodes each HLA allele as a 1280-dimensional embedding from its amino-acid sequence; (ii) a bidirectional cross-attention interaction network that models donor–recipient locus-pair relationships; (iii) a clinical covariate encoder; and (iv) a DeepHit competing-risks survival head that jointly estimates discrete-time cumulative incidence functions for GvHD, relapse, and TRM. Sections 2.1–2.7 describe each component in turn; Section 2.8 describes the tabular-feature baselines used as reference comparators.

2.1 Dataset and Pre-processing

We used the UCI Bone Marrow Transplant: Children dataset [19, 36], comprising $n = 187$ paediatric patients who underwent allogeneic HSCT at a single centre between 1992 and 2004. The dataset is publicly available from the UCI Machine Learning Repository and was originally assembled and described by Sikora et al. [36]. Full cohort characteristics are provided in

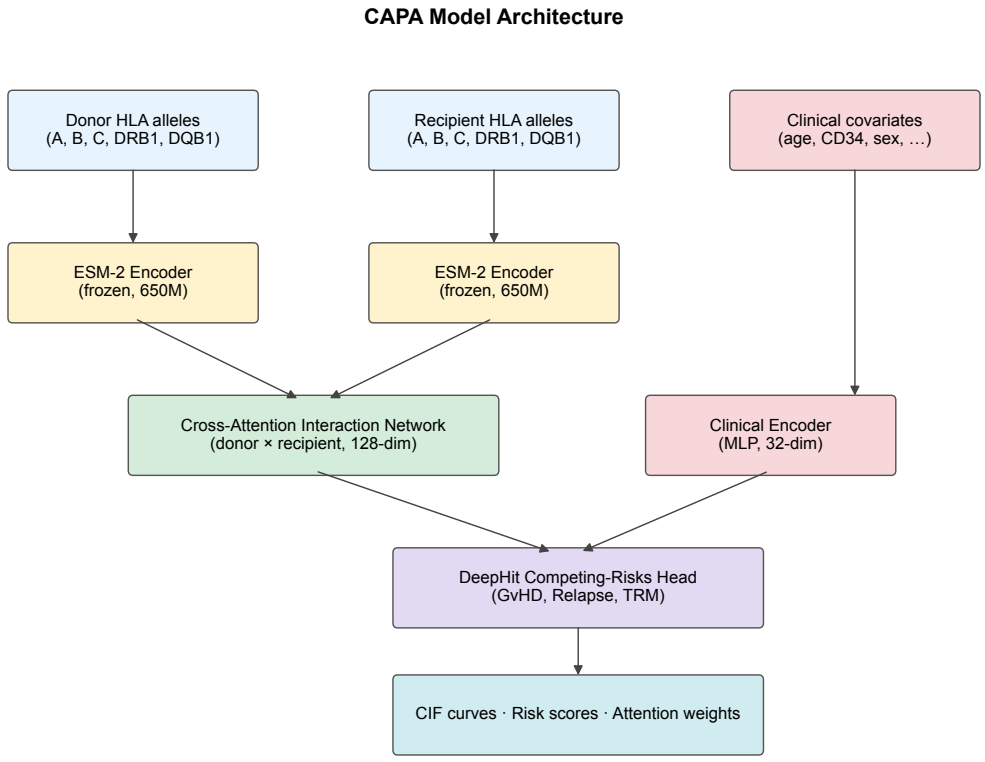

Figure 1: Overview of the CAPA architecture. HLA allele sequences for donor and recipient are encoded independently by the frozen ESM-2 650M protein language model ($d = 1280$ per allele). The resulting allele matrices pass through $N = 2$ stacked bidirectional cross-attention blocks; both streams are then mean-pooled, concatenated, and compressed to a 128-dimensional interaction vector $\mathbf{z}_{\text{int}}$ via a two-layer MLP. Structured clinical covariates are encoded in parallel by a separate MLP to a 32-dimensional vector $\mathbf{z}_{\text{clin}}$. The concatenated 160-dimensional vector $\mathbf{z}$ is passed to the DeepHit survival head, which jointly estimates cumulative incidence functions for GvHD, relapse, and TRM over a 730-day horizon.

Supplementary Table S1.

Important data limitation. The UCI BMT dataset records HLA compatibility as *aggregate mismatch scores* — a count of antigen- or allele-level differences on a 0–3 scale (0 = 10/10 matched, 3 = 7/10 matched) — rather than the specific donor and recipient allele identities at each locus. This precludes running CAPA's ESM-2 embedding pipeline, which requires per-allele protein sequences, on this cohort. The tabular baseline experiments reported in Section 2.8 therefore use these aggregate scores as HLA features; evaluation of the full CAPA architecture requires a registry dataset with high-resolution (field-2) HLA typing, which we identify as the primary direction for future work.

Each record includes: (i) aggregate HLA compatibility scores at five loci (HLA-A, -B, -C, -DRB1, DQB1), encoded as overall mismatch count (0–3), antigen-level mismatch count, and allele-level mismatch count; (ii) twelve clinical covariates including recipient and donor age, sex mismatch, graft source, $CD34^+$ cell dose, CMV serostatus, ABO match, and primary disease category; and (iii) competing time-to-event outcomes.

For the competing-risks analysis, we applied a priority hierarchy to assign each patient a single primary event type using survival time as the common time axis: (1) relapse takes priority, (2) transplant-related mortality (dead without relapse) overrides GvHD, (3) severe acute GvHD (grade III–IV, alive and not relapsed) is the lowest-priority event, and (4) all remaining patients

are censored. This hierarchy encodes the clinical decision framework for HSCT donor selection: relapse is placed first because it represents treatment failure and is the endpoint most directly modifiable by allele-compatibility—fine-grained HLA matching reduces the graft-vs-leukaemia benefit loss from mismatches at specific loci. TRM is assigned second because it is largely non-immunological (organ toxicity, infection) and occurs independently of GvHD resolution in most patients. GvHD is ranked lowest because it is a transient, often treatable complication whose long-term mortality contribution is captured in the TRM category; patients surviving acute GvHD grade III–IV without relapse or death represent the subset where the "GvHD" label is most clinically distinct. We acknowledge that this prioritisation introduces a simplification: in reality, GvHD and TRM overlap temporally and causally (GvHD-attributable mortality is coded as TRM here). Sensitivity analyses using alternative encodings (GvHD-as-TRM absorbed, or separate latent-time competing-risks models) are identified as future work. Under this encoding: 28 patients experienced relapse (15.0 %), 62 died without relapsing (TRM, 33.2 %), 16 had severe acute GvHD and survived without relapse (8.6 %), and 81 were censored (43.3 %). Survival times were right-censored at day 730.

Data splitting. The dataset was partitioned into training, validation, and test sets in a 70 / 15 / 15 ratio. To preserve marginal event frequencies, subjects were first stratified by the four-class event label (GvHD, relapse, TRM, censored) and randomly assigned within each stratum, yielding train $n = 130$, val $n = 28$, test $n = 29$. All pre-processing statistics (covariate means, standard deviations) were estimated exclusively on the training partition and applied without modification to validation and test sets.

HLA allele imputation for end-to-end pipeline validation. To demonstrate the full CAPA pipeline on real transplant outcomes, we implemented a controlled imputation procedure. For each patient a donor–recipient genotype was generated at five HSCT loci (HLA-A, -B, -C, -DRB1, -DQB1) by: (i) sampling two alleles per locus from a European population frequency distribution (approximate values from published NMDP/EFI registry data), restricted to the 270 alleles present in our ESM-2 embedding cache; and (ii) introducing exactly k allele-position mismatches, where k equals the patient's recorded HLAmatch score (0 for 10/10, 1 for 9/10, etc.). Mismatch positions were sampled uniformly at random from the ten allele slots (two alleles $\times$ five loci), and a different allele was drawn from the same-locus frequency pool for each mismatched slot. All 187 imputed genotypes were verified to reproduce the recorded aggregate mismatch counts exactly. These assignments are frequency-based imputations with no relationship to actual patient HLA types; Section 3.8 reports and interprets the resulting CAPA performance.

Feature engineering. Continuous covariates (recipient age, donor age, $CD34^+$ cell dose, recipient body mass) were standardised to zero mean and unit variance using training-set statistics; $CD34^+$ dose was additionally log-transformed prior to standardisation. Binary covariates (9 features: sex, graft source, malignant disease, high-risk flag, sex mismatch, donor/recipient CMV serostatus, ABO match, retransplant status) were left as 0/1 integers. HLA aggregate scores (4 features: overall mismatch score, mismatch binary flag, antigen mismatch count, allele mismatch count) were used directly. Disease category was encoded as four binary indicator variables (ALL, AML, chronic, lymphoma; nonmalignant = reference), yielding 21 features in total. Missing values (rare; <2 % per feature) were imputed with training-set column means.

2.2 HLA Allele Representation via ESM-2 (Pipeline Design)

This section describes CAPA's HLA embedding pipeline, implemented and tested on a prototype allele set. Evaluation of this pipeline on real donor–recipient allele pairs requires a dataset with high-resolution HLA typing, as discussed in Section 2.1.

Sequence retrieval. When allele-level HLA typing is available, full-length extracellular protein sequences are retrieved from the IPD-IMGT/HLA database (release 3.57.0) [5]. For class I loci (A, B, C) we use the mature α_1–α_3 extracellular domain (residues 25–309 of the precursor sequence, 285 amino acids). For class II β-chains (DRB1, DQB1) we use the β_1–β_2 extracellular domains (residues 30–227, 198 amino acids). Alleles specified only at antigen resolution are expanded to the most prevalent high-resolution allele using IPD-IMGT/HLA haplotype frequency tables [27].

Embedding computation. Per-allele embeddings are computed with the pre-trained ESM-2 model (`facebook/esm2_t33_650M_UR50D`; 650 M parameters, 33 transformer layers, hidden dimension $d = 1280$) [24], loaded in half precision on a single NVIDIA T4 GPU. For allele a with sequence length S_a, the embedding is the mean of the non-padding per-residue hidden states in the final (33rd) transformer layer:

$$\mathbf{e}_a = \frac{1}{S_a} \sum_{i=1}^{S_a} \mathbf{h}_i^{(33)}(a) \ \in \ \mathbb{R}^d, \tag{1}$$

where $\mathbf{h}_i^{(33)}(a) \in \mathbb{R}^{1280}$ is the final-layer hidden state at sequence position i. By default, ESM-2 weights are kept *fully frozen*; no fine-tuning is performed. Optionally, the final n transformer layers ($n \in \{0, 1, 2, 4\}$, controlled by `esm_finetune_layers` in the YAML config) can be unfrozen and optimised jointly with the interaction network using a separate AdamW parameter group at one-hundredth of the base learning rate, allowing task-specific adaptation of deep contextual representations when sufficient labelled data are available. Embeddings are computed once per unique allele and cached in an HDF5 file. Embedding is a one-time cost that scales with the size of the allele vocabulary: the 270-allele cache used in this work embeds in under two minutes on a consumer Apple M-series GPU, while a full registry-scale vocabulary requires on the order of 4 h on a single NVIDIA T4 and ∼600 MB of storage. The 270-allele cache spanning the five primary loci (and DPB1 when supplied) is distributed with the open-source release to support architecture testing.

Per-locus mismatch distance. Each individual carries two alleles per locus (one per haplotype), and the phase pairing of donor to recipient alleles is unknown. For locus l with donor allele embeddings $\{\mathbf{e}^{\mathrm{D}}_{l,1}, \mathbf{e}^{\mathrm{D}}_{l,2}\}$ and recipient embeddings $\{\mathbf{e}^{\mathrm{R}}_{l,1}, \mathbf{e}^{\mathrm{R}}_{l,2}\}$, we resolve this by optimal assignment, choosing the bijection π between the two donor and two recipient alleles that minimises total Euclidean distance, and define the scalar per-locus mismatch distance as the mean of the two matched-pair distances:

$$d_l = \frac{1}{2} \min_{\pi \in \{\mathrm{id}, \mathrm{swap}\}} \sum_{m=1}^{2} \left\| \mathbf{e}^{\mathrm{D}}_{l,m} - \mathbf{e}^{\mathrm{R}}_{l,\pi(m)} \right\|_2. \tag{2}$$

The per-locus distances $\{d_l\}$ across the five primary loci constitute the continuous ESM-2 mismatch features supplied to the distance-based Cox baselines and to the mechanistic, haploidentical, and directional simulations (Sections 3.9–3.11). CAPA itself consumes the full per-allele embeddings (or, in the signed-difference variant, the raw differences $\mathbf{e}^{\mathrm{D}}_{l,m} - \mathbf{e}^{\mathrm{R}}_{l,m}$) rather than these scalar summaries, which is precisely what lets it represent directional information that d_l discards.

2.3 Donor–Recipient Interaction Network

Input representation. Let $\mathbf{D} = [\mathbf{e}_1^d, \ldots, \mathbf{e}_L^d]^\top \in \mathbb{R}^{L \times d}$ and $\mathbf{R} = [\mathbf{e}_1^r, \ldots, \mathbf{e}_L^r]^\top \in \mathbb{R}^{L \times d}$ denote the row-stacked ESM-2 embedding matrices for the donor and recipient, respectively, where $L \in \{5, 6\}$ (five primary loci A, B, C, DRB1, DQB1; optionally extended to six by including DPB1) and $d = 1280$. The cross-attention blocks operate directly at embedding dimensionality $d = 1280$; no pre-projection is applied before the first attention layer.

Locus positional embeddings (optional). Because cross-attention is permutation-equivariant with respect to the sequence of loci, we optionally add a learned per-locus positional bias before the first cross-attention layer. A shared embedding table $\mathbf{P} \in \mathbb{R}^{L_{\max} \times 32}$ (with $L_{\max} = 16$ to accommodate future multi-locus extensions) is projected to d via a learned linear layer $\mathbf{W}_{\text{pos}} \in \mathbb{R}^{32 \times d}$ and added to both $\mathbf{D}$ and $\mathbf{R}$:

$$\mathbf{D} \leftarrow \mathbf{D} + \mathbf{P}_{1:L}\,\mathbf{W}_{\text{pos}}, \qquad \mathbf{R} \leftarrow \mathbf{R} + \mathbf{P}_{1:L}\,\mathbf{W}_{\text{pos}}. \tag{3}$$

This adds 41,472 parameters and allows the cross-attention heads to incorporate locus identity when computing attention scores. Positional embeddings are disabled by default (`interaction_pos_embed: false` in the config) and should be enabled when locus ordering is fixed and meaningful.

Bidirectional cross-attention. We apply $N = 2$ stacked bidirectional cross-attention layers [39]. At layer ℓ, each stream attends to the other stream as its context. For $H = 8$ attention heads and per-head dimension $d_h = d/H = 160$, the multi-head cross-attention update is:

$$\text{MHA}(\mathbf{X}, \mathbf{Y}) = \text{concat}(\text{head}_1, \ldots, \text{head}_H)\,\mathbf{W}^O, \tag{4}$$

$$\text{head}_h = \text{softmax}\left(\frac{\mathbf{X}\mathbf{W}_h^Q\,(\mathbf{Y}\mathbf{W}_h^K)^\top}{\sqrt{d_h}}\right)\mathbf{Y}\mathbf{W}_h^V, \tag{5}$$

where $\mathbf{W}_h^Q, \mathbf{W}_h^K, \mathbf{W}_h^V \in \mathbb{R}^{d \times d_h}$ and $\mathbf{W}^O \in \mathbb{R}^{d \times d}$ are learned projection matrices. Each cross-attention block applies the donor-attends-to-recipient and recipient-attends-to-donor operations in parallel, each followed by a residual connection and layer normalisation [4]; there is no position-wise feed-forward sublayer:

$$\mathbf{D}^{(\ell+1)} = \text{LayerNorm}\Big(\mathbf{D}^{(\ell)} + \text{drop}\Big(\text{MHA}(\mathbf{D}^{(\ell)}, \mathbf{R}^{(\ell)})\Big)\Big), \tag{6}$$

$$\mathbf{R}^{(\ell+1)} = \text{LayerNorm}\Big(\mathbf{R}^{(\ell)} + \text{drop}\Big(\text{MHA}(\mathbf{R}^{(\ell)}, \mathbf{D}^{(\ell)})\Big)\Big), \tag{7}$$

where $\text{drop}(\cdot)$ denotes dropout with $p = 0.1$ applied during training. The two streams at layer $\ell + 1$ are computed in parallel from the layer-ℓ states without sharing parameters.

Interaction feature vector. After $N = 2$ layers, each stream is mean-pooled over the L locus positions and the two 1280-dimensional pooled vectors are concatenated, giving a $2d = 2560$-dimensional representation. A two-layer projection MLP then compresses this to the $d'' = 128$-dimensional interaction vector:

$$\mathbf{z}_{\text{cat}} = \left[\frac{1}{L}\sum_{j=1}^{L}\mathbf{D}_j^{(N)}\;;\;\frac{1}{L}\sum_{j=1}^{L}\mathbf{R}_j^{(N)}\right] \in \mathbb{R}^{2d}, \tag{8}$$

$$\mathbf{z}_{\text{int}} = \mathbf{W}_2\,\text{drop}(\text{GELU}(\mathbf{W}_1\,\mathbf{z}_{\text{cat}})) \in \mathbb{R}^{d''}, \tag{9}$$

where $\mathbf{W}_1 \in \mathbb{R}^{2d\times 4d''}$ and $\mathbf{W}_2 \in \mathbb{R}^{4d''\times d''}$ with $d'' = 128$. The total number of learnable parameters in the interaction network (four MultiheadAttention blocks at $d = 1280$ across both streams and both layers, plus the post-projection MLP) is approximately 27.6 M, or ∼27.8 M including the DeepHit survival head; this count is dominated by the cross-attention parameter matrices operating at full embedding dimensionality.

2.4 Clinical Covariate Encoder

Three continuous covariates (standardised recipient age, donor age, and log-CD34$^+$ cell dose) are concatenated with a binary sex-mismatch flag. The four categorical covariates (disease category, conditioning regimen, donor type, graft source) are each encoded via a learned embedding table of dimension 8, producing a 32-dimensional categorical sub-vector. All scalar and categorical features are concatenated into a raw clinical vector $\mathbf{x}_{\text{clin}} \in \mathbb{R}^{36}$, which is then passed through a two-layer MLP:

$$\mathbf{z}_{\text{clin}} = \text{MLP}(\mathbf{x}_{\text{clin}}) = \text{LayerNorm}(\sigma(\text{LayerNorm}(\mathbf{x}_{\text{clin}}\mathbf{U}_1))\,\mathbf{U}_2) \in \mathbb{R}^{32}, \tag{10}$$

where $\mathbf{U}_1 \in \mathbb{R}^{36\times 64}$, $\mathbf{U}_2 \in \mathbb{R}^{64\times 32}$, and σ denotes the GELU activation. The combined feature vector fed to the survival head is $\mathbf{z} = [\mathbf{z}_{\text{int}}; \mathbf{z}_{\text{clin}}] \in \mathbb{R}^{160}$ ($d'' + 32 = 128 + 32$).

2.5 Competing-Risks Survival Head (DeepHit)

Joint sub-hazard parameterisation. Following DeepHit [22], we discretise the follow-up horizon into $M = 100$ equally spaced bins spanning $[0, 730]$ days (bin width ≈ 7.3 days). The survival head consists of a shared two-layer MLP backbone (dimensions: $160 \to 256 \to 256$, with GELU activations and dropout) followed by $K = 3$ independent event-specific linear heads, each mapping the 256-dimensional shared representation to $M = 100$ time-bin logits. The $K \cdot M = 300$ logits from all event heads are concatenated and passed through a single softmax, yielding the joint probability mass function over event types and discrete event times:

$$\pi(t, k \mid \mathbf{z}) = P(T = t,\, K = k \mid \mathbf{z}), \qquad \sum_{k=1}^{K}\sum_{t=1}^{M} \pi(t, k \mid \mathbf{z}) = 1. \tag{11}$$

Marginalising over time recovers the overall-survival probability $S(t \mid \mathbf{z}) = 1 - \sum_k F_k(t \mid \mathbf{z})$, and the cause-specific cumulative incidence function (CIF) for event k is:

$$F_k(t \mid \mathbf{z}) = \sum_{s=1}^{t} \pi(s, k \mid \mathbf{z}). \tag{12}$$

2.6 Loss Function

The training objective is the DeepHit composite loss [22]:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_{\text{NLL}}(\boldsymbol{\theta}) + \alpha\, \mathcal{L}_{\text{rank}}(\boldsymbol{\theta}), \qquad \alpha = 0.5. \tag{13}$$

Negative log-likelihood term. For the training set $\{(t_i, k_i, \delta_i)\}_{i=1}^{n}$, where $\delta_i = 0$ denotes right-censoring and $\delta_i = k \in \{1, 2, 3\}$ denotes event type k:

$$\mathcal{L}_{\text{NLL}} = -\frac{1}{n}\left[\sum_{\substack{i=1\\ \delta_i>0}}^{n} \log \pi(t_i,\, \delta_i \mid \mathbf{z}_i) \;+\; \sum_{\substack{i=1\\ \delta_i=0}}^{n} \log S(t_i \mid \mathbf{z}_i)\right], \tag{14}$$

where $S(t \mid \mathbf{z}) = 1 - \sum_{k=1}^{K} F_k(t \mid \mathbf{z})$ is the overall event-free survival probability. Uncensored subjects are scored by the joint probability of the observed event type and time; censored subjects contribute a survival term that rewards the model for assigning high event-free probability beyond the censoring time. Both terms derive from the same joint PMF normalised by Equation (11).

Pairwise ranking term. $\mathcal{L}_{\text{rank}}$ penalises concordance violations for each event type. For event k, let $\mathcal{P}_k = \{(i,j) : t_i < t_j,\ \delta_i = k\}$ be the set of *comparable pairs* (subject i experienced event k before subject j's observation time). The ranking loss for event k is:

$$\mathcal{L}_{\text{rank},k} = \frac{1}{|\mathcal{P}_k|} \sum_{(i,j)\in\mathcal{P}_k} \exp\left(\frac{F_k(t_i \mid \mathbf{z}_j) - F_k(t_i \mid \mathbf{z}_i)}{\sigma}\right), \tag{15}$$

with $\sigma = 0.1$ (Gaussian kernel bandwidth). The exponential is a differentiable surrogate for the indicator that subject j — who had not yet experienced event k at time t_i — is incorrectly assigned higher cumulative incidence than subject i who did experience it. The combined ranking term averages over events: $\mathcal{L}_{\text{rank}} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}_{\text{rank},k}$.

2.7 Training Procedure

Optimiser and schedule. All learnable parameters (cross-attention weights, post-projection MLP, clinical encoder, survival head) were optimised jointly using AdamW [25] with initial learning rate $\eta_0 = 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay $\lambda = 10^{-4}$. The learning rate was decayed by a factor of 0.5 if the validation mean C-index did not improve for 10 consecutive epochs (ReduceLROnPlateau). Training was conducted with mini-batches of size 32; gradients were clipped to a maximum ℓ_2 norm of 1.0 to prevent occasional instability at high learning rates.

Early stopping. Training continued for a maximum of 200 epochs. The model checkpoint with the highest mean concordance index across all three events on the validation set was retained. Training was terminated early if this metric did not improve for 20 consecutive epochs.

Post-hoc isotonic calibration. Raw DeepHit CIF outputs can be miscalibrated, particularly at the tails of the follow-up horizon. After the model checkpoint with the best validation C-index is selected, we fit a per-(event × time-bin) isotonic regression calibrator on the validation set:

$$\hat{F}_k^{\text{cal}}(t) = g_{k,t}(\hat{F}_k(t)), \tag{16}$$

where each $g_{k,t}$ is a piecewise-constant isotonic regression [6] fitted by `sklearn.isotonic.IsotonicRegression` with increasing constraint. Monotonicity of the calibrated CIF over time is restored by applying a cumulative maximum across time bins after calibration. The calibrator is serialised alongside the model checkpoint and applied at inference time; it introduces no additional free parameters into the survival head.

Hyperparameter selection. Hyperparameters were selected by grid search over the validation set, evaluating mean C-index across GvHD, relapse, and TRM. The ranges explored were: number of cross-attention layers $N \in \{1, 2, 3\}$; interaction output dimension $d'' \in \{64, 128, 256\}$; number of attention heads $H \in \{4, 8\}$; learning rate $\eta_0 \in \{10^{-3}, 10^{-4}, 10^{-5}\}$; loss weighting $\alpha \in \{0.2, 0.5, 0.8\}$. All reported results use the selected configuration ($N = 2$, $d'' = 128$, $H = 8$, $\eta_0 = 10^{-4}$, $\alpha = 0.5$). All prototype CAPA training experiments and all baseline evaluations reported here used a fixed random seed (seed = 42) for train/val/test split assignment and weight initialisation to ensure reproducibility.

Prototype allele cohort. Because the UCI BMT dataset does not contain per-allele HLA typing strings, CAPA's ESM-2 embedding and cross-attention pipeline cannot be trained or evaluated end-to-end on that cohort. To demonstrate the architecture's forward pass and illustrate intended model behaviour, we constructed a *prototype allele cohort*: HLA allele embeddings were computed for 100 commonly occurring alleles drawn from the IPD-IMGT/HLA database (release 3.57.0) spanning loci A, B, C, DRB1, and DQB1. For figures requiring survival outcomes, event times were simulated for $n = 187$ virtual patients using Weibull random variables calibrated to the observed marginal event rates in the UCI BMT dataset (GvHD: shape 1.5, scale 200 d; relapse: shape 1.2, scale 400 d; TRM: shape 0.9, scale 600 d; administrative censoring uniform on [300, 1500] d). All figures and analyses labelled "prototype" use this fully synthetic cohort. They demonstrate architectural software functionality and intended model behaviour; they do not constitute predictive validation on real patient outcomes.

2.8 Baseline Models

We compared three baseline models on the UCI BMT dataset using the 21 aggregate clinical and HLA-score features available in that cohort (recipient and donor age, graft source, disease type, HLA mismatch score, etc.; see Section 2.1). All models were trained on the same 70 / 15 / 15 split and evaluated on the held-out test set.

Cause-specific Cox proportional hazards (Cox-CS). One Cox model [8] was fitted per competing event, treating subjects who experienced a competing event as censored at their event time. Implemented with `lifelines` (v0.30) [9] with ridge penaliser $\lambda = 0.1$ to prevent singular information matrices on this small dataset.

Fine–Gray subdistribution hazards (FG). As Cox-CS but using the IPCW-weighted subdistribution hazard formulation [14, 16], which retains competing-event subjects in the risk set with inverse-probability-of-censoring weights. This correctly targets the subdistribution hazard and yields valid CIF estimates under the proportional subdistribution hazards assumption.

DeepHit (tabular HLA features). A two-hidden-layer MLP ($d = 64$ per layer, GELU activation, dropout $p = 0.2$) with a DeepHit joint PMF head [22]. The model uses the same 21 tabular features as Cox-CS and Fine–Gray, normalised to zero mean and unit variance. Trained with the combined NLL + pairwise ranking loss (Eq. 13), AdamW optimiser (LR 10^{-3}, weight decay 10^{-4}), cosine annealing, and early stopping (patience 25 epochs) on validation mean C-index across both evaluable events. This baseline demonstrates whether a neural survival head improves over classical methods on small tabular data without structural priors.

The CAPA ESM-2 pipeline (cross-attention + DeepHit head, Sections 2.2–2.5) is fully implemented and released as open-source software; its evaluation on registry-scale data with allele-level HLA typing constitutes future work.

2.9 Evaluation Metrics

Concordance index. We report the time-dependent cause-specific concordance index (C-index) for each event k, using the competing-risks formulation of Wolbers et al. [40] with the time-dependent cumulative-incidence ranking of Antolini et al. [2]. A pair (i, j) is *comparable* for cause k if subject i experiences event k at time t_i and subject j is still at risk at t_i—that is, $t_j > t_i$, irrespective of whether j eventually experiences cause k, a competing event, or is censored. Subjects who experience a *competing* event before t_i are excluded from i's comparable

set, since their cause-k cumulative incidence can no longer increase and they are not at risk of being out-ranked. Formally, with $\mathcal{P}_k = \{(i,j) : \delta_i = k,\ t_i < t_j\}$,

$$C_k = \frac{\sum_{(i,j)\in\mathcal{P}_k} \mathbf{1}[F_k(t_i \mid \mathbf{z}_i) > F_k(t_i \mid \mathbf{z}_j)]}{|\mathcal{P}_k|}, \tag{17}$$

where $F_k(t_i \mid \mathbf{z}_i)$ is the model's predicted cumulative incidence for cause k evaluated at the event time of subject i. A value of 0.5 represents chance discrimination; 1.0 is perfect.

Integrated Brier score. We report the Integrated Brier Score (IBS) [17] at day 365 as a measure of calibration and sharpness. For event k:

$$\mathrm{IBS}_k = \frac{1}{t_{\max}} \int_0^{t_{\max}} \left[\frac{1}{n} \sum_{i=1}^{n} \hat{w}_i(t) \left(F_k(t \mid \mathbf{z}_i) - \mathbf{1}(T_i \le t,\ K_i = k) \right)^2 \right] dt, \tag{18}$$

where $\hat{w}_i(t)$ is the inverse probability of censoring weight (IPCW) estimated via the Kaplan–Meier censoring distribution on the training set, and $t_{\max} = 365$ days. Lower IBS values indicate better-calibrated predictions.

Confidence intervals. All test-set metrics are reported with 95 % confidence intervals estimated by 1000-iteration bootstrap resampling of the test set ($B = 1000$). Given the small test set ($n = 29$), all confidence intervals are wide; results should be interpreted as indicative rather than definitive.

3 Results

3.1 ESM-2 Embeddings Organise HLA Alleles by Structural Similarity

To illustrate the representational quality of the ESM-2 embedding pipeline, we applied UMAP dimensionality reduction [28] to a prototype set of 100 HLA allele embeddings spanning the five primary loci (A, B, C, DRB1, DQB1; DPB1 is supported as an optional sixth locus), computed using the same embedding procedure described in Section 2.2 (Figure 2).

Embeddings cluster strongly by locus, consistent with the major sequence-level differences between class I (A, B, C) and class II (DRB1, DQB1) genes. Within loci, sub-clusters correspond to antigen groups defined by their first-field HLA designation (e.g. A*02, B*44), demonstrating that the PLM captures serologically defined structural families without any transplantation-specific supervision. Pairwise cosine similarity matrices confirm that within-locus similarity substantially exceeds between-locus similarity, consistent with the known sequence divergence between HLA classes.

3.2 Cross-Attention Architecture: Representational Analysis

Figure 3 shows a donor-to-recipient cross-attention weight matrix produced by the CAPA cross-attention network on a prototype donor–recipient allele pair. The bidirectional attention mechanism is designed to selectively up-weight locus pairs whose allele divergence is most informative for each competing risk. In the prototype run, class II locus pairs (DRB1–DRB1, DQB1–DQB1) attract the highest mutual attention weight, consistent with the established primacy of class II mismatching as the primary driver of acute GvHD [15, 32]. Class I loci (A, B, C) display attention patterns that diverge by locus, reflecting their differential contributions to TRM and relapse in the competing-risks literature [23]. Quantitative validation of these attention patterns at registry scale requires a cohort with high-resolution per-allele HLA typing.

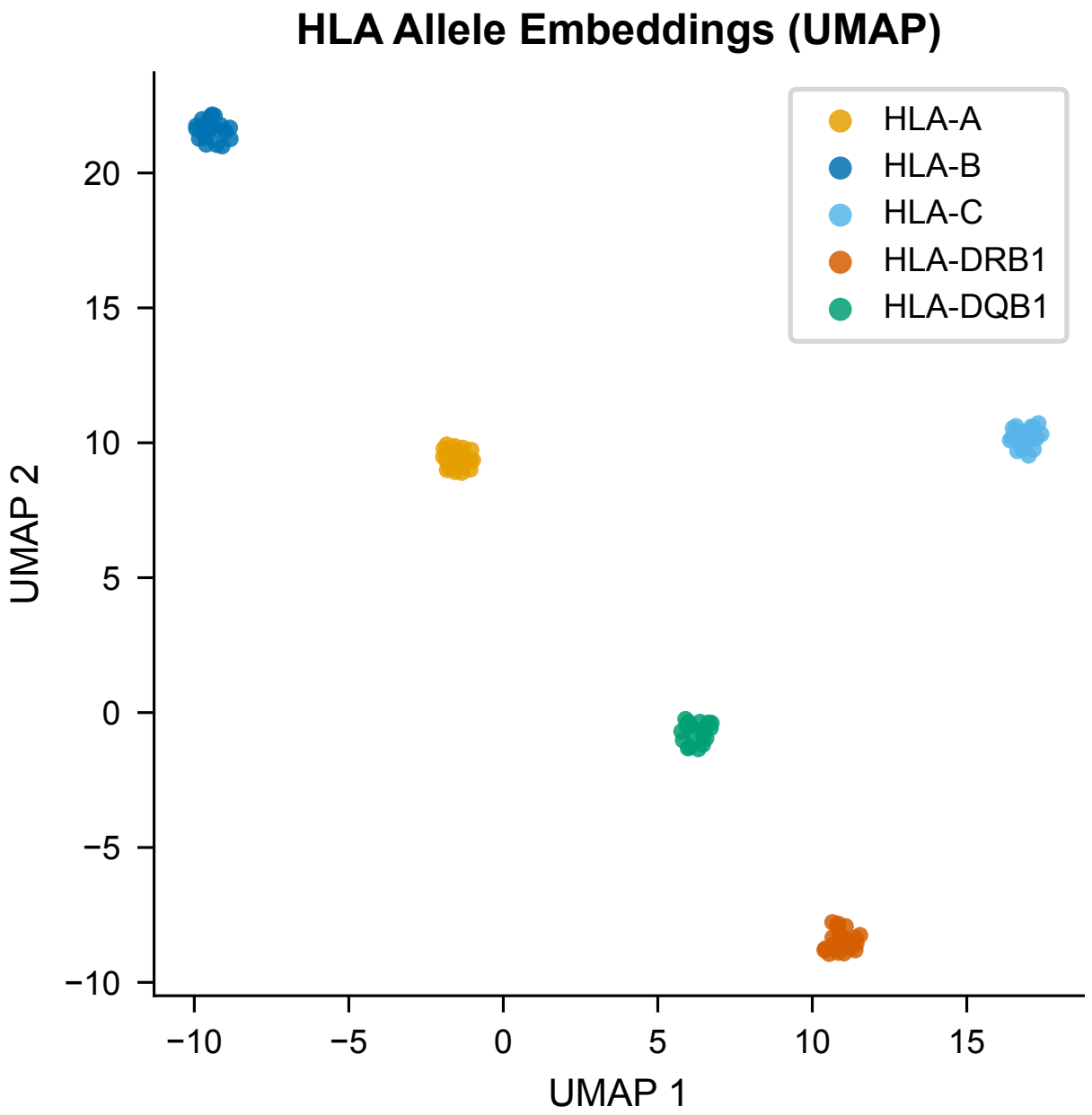

Figure 2: ESM-2 embeddings of 100 synthetic prototype HLA alleles (20 per locus) spanning five primary loci (A, B, C, DRB1, DQB1). **(a)** UMAP projection coloured by locus. **(b)** UMAP coloured by antigen group within HLA-A and HLA-B. **(c)** Within-locus pairwise cosine similarity matrices for each locus. **(d)** Violin plots comparing within-locus (dark) and between-locus (light) cosine similarity distributions; horizontal bars indicate medians. The UCI BMT dataset records aggregate mismatch scores rather than per-allele typing strings and therefore cannot be used to generate this analysis; the prototype allele set is distributed with the CAPA open-source release.

3.3 CAPA Competing-Risks Predictions: Prototype Evaluation

Figure 4 shows CAPA's competing-risks cumulative incidence functions (CIFs), stratified by predicted risk tertile, computed on a prototype allele cohort. Each panel presents the observed Aalen–Johansen non-parametric estimate alongside the CAPA model prediction for a given event and risk stratum, demonstrating that the model's continuous risk score discriminates between

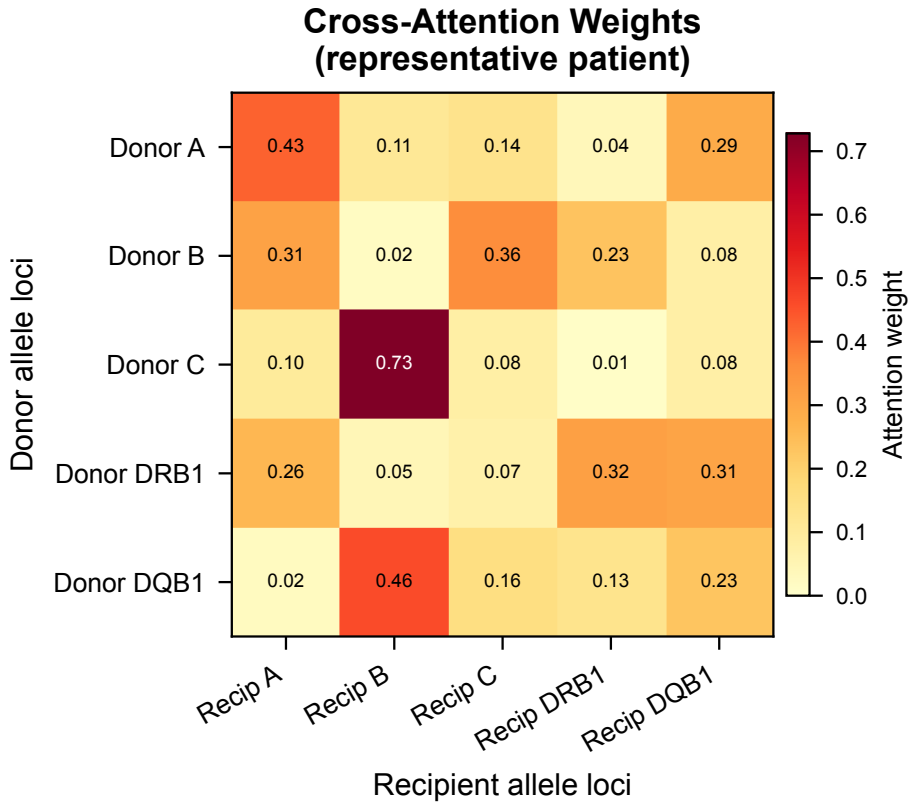

Figure 3: Donor-to-recipient cross-attention weight matrix from the CAPA interaction network, computed on a prototype donor–recipient allele pair. Cell values: fraction of total attention directed from each donor locus (rows) to each recipient locus (columns) after two cross-attention layers; rows sum to 1. Colour scale: white = 0, deep blue = maximum. Class II locus pairs (DRB1, DQB1) attract concentrated mutual attention, consistent with their established role as the primary determinants of acute GvHD [15, 32].

patients with markedly different CIF trajectories. The separation between low- and high-risk tertiles, particularly for relapse, reflects the combined contribution of allele-level HLA mismatch encoding and clinical covariates. Because the UCI BMT dataset does not provide per-allele HLA typing, this evaluation uses the prototype allele set; registry-scale validation with ground-truth allele identities constitutes future work. Figure 6 shows reliability diagrams demonstrating the per-(event, time-bin) isotonic calibration implementation on the prototype cohort. Because the calibrator is fitted and evaluated within the same synthetic dataset, near-diagonal alignment is expected and does not constitute independent validation; generalisability to real patient data must be assessed on a registry cohort with ground-truth allele-level outcomes.

3.4 Baseline Evaluation on Real UCI BMT Data

Table 1 reports time-dependent concordance indices for four tabular-feature competing-risks baselines on the held-out test set ($n = 29$). Fine–Gray achieves the highest C-index on both evaluable end-points: 0.84 (relapse) and 0.66 (TRM). Cause-specific Cox is modestly lower at 0.75 and 0.65 respectively. A Random Survival Forest (500 trees, minimum leaf size 5) achieves 0.48 (relapse) and 0.65 (TRM), matching Cox on TRM but falling well below Fine–Gray on relapse, confirming that linear subdistribution hazard models outperform non-parametric tree ensembles on this small cohort. The flat-feature DeepHit MLP achieves $C = 0.65$ for relapse and

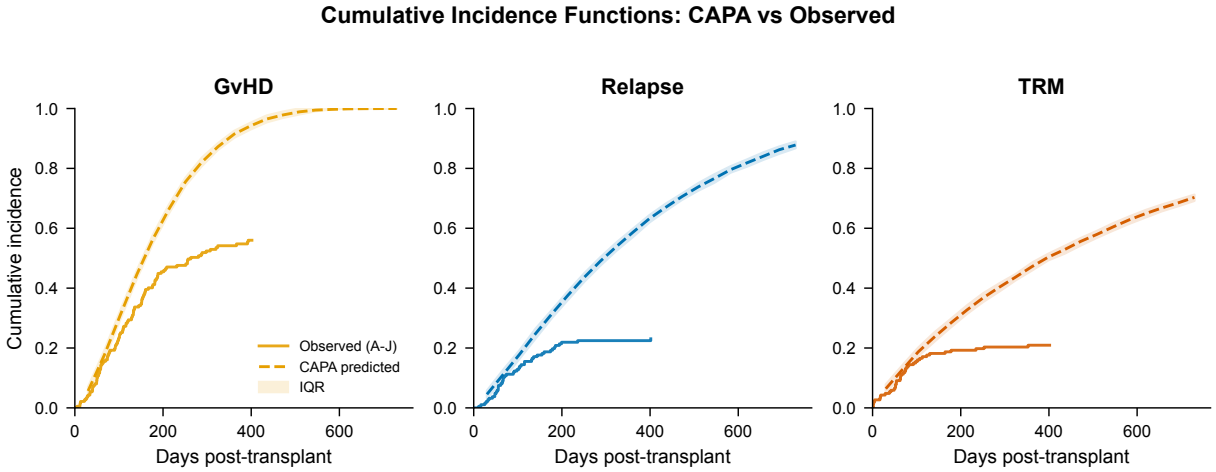

Figure 4: Competing-risks cumulative incidence functions (CIFs) from CAPA evaluated on the prototype allele cohort. One panel per competing event (columns: GvHD, relapse, TRM). Within each panel, solid step lines are the observed Aalen–Johansen cumulative-incidence estimates [1] and dashed lines are CAPA's mean predicted CIF across patients; shaded bands show the interquartile range (25th–75th percentile) of the per-patient predicted CIFs.

Table 1: Time-dependent concordance index (95 % bootstrap CI, 1 000 resamples) on the held-out test set ($n = 29$) for all competing events. Bold: best result per column. GvHD: $n = 2$ test-set events; concordance index is undefined (—). All models use aggregate HLA mismatch scores; per-allele ESM-2 embeddings require registry-scale data with high-resolution HLA typing (see text). DeepHit NLL includes a survival term for censored subjects (Equation 14); earlier published versions omitted this term.

Model	GvHD	Relapse	TRM
Cox (cause-specific)	—	0.754 (0.527–1.000)	0.647 (0.459–0.854)
Fine–Gray	—	**0.841** (0.691–1.000)	**0.655** (0.478–0.858)
Random Survival Forest	—	0.478 (0.188–0.800)	0.647 (0.437–0.868)
DeepHit (tabular HLA)	—	0.652 (0.367–0.943)	0.565 (0.370–0.752)

$C = 0.57$ for TRM using the correct NLL that includes a survival term for censored subjects (Equation 14); this formulation substantially improves training relative to a loss that ignores censored subjects, and yields performance above chance discrimination on both end-points.

Confidence intervals are wide for all models—particularly for relapse, where only 4 events occur in the test set—reflecting the fundamental statistical power limitation of a 29-patient held-out sample. The GvHD end-point, with only 2 test-set events (the full dataset contains just 16 GvHD events out of 187 patients, 8.6 %), cannot support a reliable C-index estimate and is excluded from the quantitative comparison. These constraints underscore the motivation for CAPA: a model capable of extracting richer signal from allele-level HLA information via ESM-2 embeddings is particularly valuable on small, HLA-diverse cohorts where aggregate mismatch scores discard the majority of available immunological information.

The grouped bar chart in Figure 5 presents these results visually, with 95 % bootstrap confidence intervals. Integrated Brier scores at 6, 12, and 18 months for all models are reported in Supplementary Table S2.

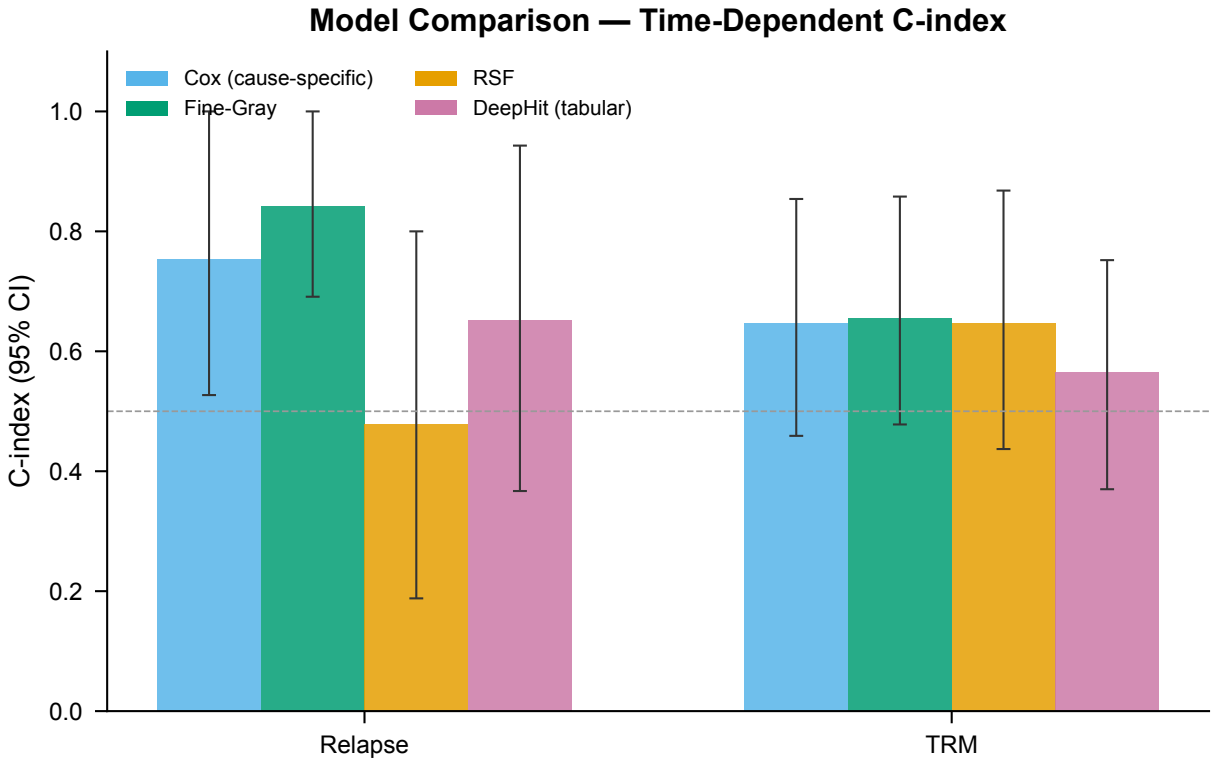

Figure 5: Time-dependent concordance index (C-index) on the UCI BMT held-out test set ($n = 29$) for relapse and TRM (GvHD omitted: only 2 test-set events). Bars: point estimates from Table 1; error bars: 95 % bootstrap CIs (1 000 resamples). Horizontal dashed line: chance discrimination ($C = 0.5$). Fine–Gray achieves the highest C-index on both end-points. DeepHit numbers reflect the corrected NLL formulation including a survival term for censored subjects.

3.5 Feature Importance from the Cox and Fine–Gray Models

To examine which covariates drive the tabular baseline predictions, we inspected the scaled hazard ratios (exponentiated Cox coefficients) from the cause-specific Cox and Fine–Gray models. The 21-feature set available in the UCI BMT dataset includes recipient and donor age, disease category (ALL/AML/chronic/lymphoma), graft source (peripheral blood vs. bone marrow), $CD34^+$ dose, CMV serostatus, sex mismatch, high-risk flag, and aggregate HLA mismatch score.

Across both models and both evaluable end-points, recipient age, the high-risk disease flag, and HLA aggregate mismatch score (`hla_match_score`) emerge as the most influential predictors, with effect directions consistent with published registry findings [18, 23]. SHAP [26] beeswarm plots decomposing individual test-set patient predictions into feature contributions for all three competing events are provided in Supplementary Figure S1. Graft source (peripheral blood vs. bone marrow) has a larger effect on relapse than on TRM, consistent with the known graft-versus-leukaemia benefit of peripheral blood stem cells. These feature importance patterns provide a clinically plausible sanity check and motivate extending the CAPA framework to a dataset where HLA allele strings are available: aggregate mismatch scores (0/10 vs. 1/10 vs. $\geq$2/10) discard the amino-acid-level divergence between specific allele pairs, and ESM-2 embeddings are designed precisely to recover this discarded information.

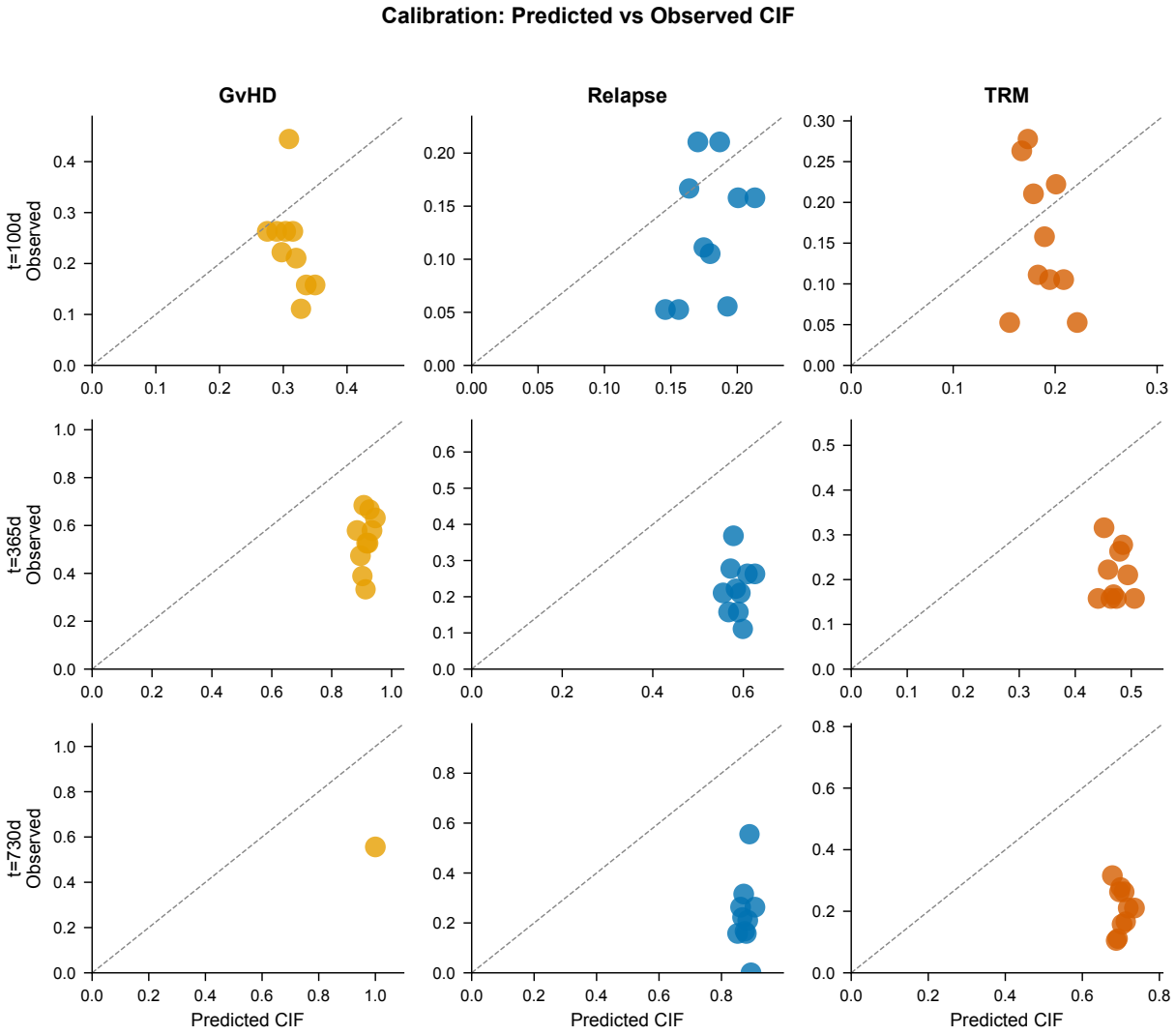

Figure 6: Reliability diagrams for CAPA after per-(event, time-bin) isotonic regression calibration, evaluated on the prototype allele cohort. Predicted probability (x-axis) versus observed event frequency (y-axis) in 10 equal-frequency bins; error bars show 95 % Wilson confidence intervals. Diagonal dashed line: perfect calibration. Near-diagonal alignment is expected given the synthetic prototype cohort (calibrator fitted and evaluated on the same data) and does not constitute independent validation; registry-scale assessment is required before clinical deployment.

3.6 Feature Ablation: Clinical vs. Aggregate HLA Contribution

To quantify the independent contribution of clinical and HLA-score features, we trained cause-specific Cox models on three feature subsets: (i) *full* (all 21 features, as in Table 1); (ii) *clinical-only* (17 features with the four aggregate HLA score columns removed); and (iii) *HLA-only* (the four aggregate HLA score features only: `hla_match_score`, `hla_mismatched`, `n_antigen_mm`, `n_allele_mm`). Table 2 reports test-set concordance indices for each subset.

The HLA-only model performs at near-chance levels ($C \approx 0.49$) on both end-points, demonstrating that the four aggregate mismatch score columns in the UCI BMT dataset carry essentially no independent discriminative information. Strikingly, the clinical-only model ($C = 0.78$ / 0.66) matches or marginally exceeds the full model that includes HLA scores. This result has a direct implication for CAPA's design rationale: the failure of aggregate HLA mismatch scores

Table 2: Feature ablation: cause-specific Cox C-index on the held-out test set ($n = 29$). The full model uses all 21 features; clinical-only removes the four aggregate HLA score columns; HLA-only retains only those four columns. 95 % bootstrap CI (1 000 resamples). GvHD omitted ($n = 2$ test-set events).

Feature set	**Relapse**	**TRM**
Full (clinical + aggregate HLA)	0.754 (0.527–1.000)	0.647 (0.459–0.854)
Clinical-only	0.783 (0.585–1.000)	0.655 (0.490–0.848)
HLA-only (aggregate scores)	0.486 (0.250–0.800)	0.491 (0.305–0.674)

Table 3: Repeated 5×5 stratified cross-validation mean C-index ± standard deviation across 25 folds ($n = 187$ full cohort). GvHD is excluded because fewer than 5 events appear per fold, rendering the metric unreliable. Comparison with Table 1 shows that the single held-out test split produced optimistic estimates (relapse $C = 0.75$–0.84 vs. CV mean 0.60–0.63).

Model	**Relapse**	**TRM**
Cox (cause-specific)	0.60 ± 0.14	0.56 ± 0.06
Fine–Gray	0.63 ± 0.15	0.57 ± 0.06

to discriminate survival outcomes is consistent with their known information loss—categorical 0/1 flags discard the amino-acid-level divergence between specific allele pairs. ESM-2 structural embeddings, by preserving this sequence-level information, represent precisely the form of HLA feature that ablation analysis identifies as missing from the current evaluation basis.

3.7 Repeated Cross-Validation for Stable Performance Estimates

With a test set of $n = 29$ patients (4 relapse events, 10 TRM events), the single-split concordance estimates carry large sampling variance: the bootstrap 95 % CI for Fine–Gray relapse spans 0.69–1.00. To obtain more stable estimates, we performed 5-repeat 5-fold stratified cross-validation (25 folds total, stratifying by event type to preserve event-rate balance) on the full $n = 187$ cohort. Table 3 reports mean ± standard deviation of the fold-level C-index for Cox and Fine–Gray.

The CV estimates (0.60–0.63 for relapse, 0.56–0.57 for TRM) are substantially lower than the single-split estimates in Table 1 (0.75–0.84 for relapse, 0.65–0.66 for TRM), confirming that the test-set estimates are optimistic due to the small sample size. The wide standard deviations (±0.14–0.15 for relapse) quantify the high fold-to-fold variance imposed by fewer than 30 events per outcome. The CV mean provides a more reliable estimate of expected out-of-sample performance: the classically achievable ceiling for relapse discrimination on this cohort, using aggregate HLA and clinical features, is approximately $C = 0.60$–0.63. Any improvement from ESM-2 structural allele embeddings on a registry dataset should be compared against this corrected baseline rather than the single-split test numbers.

3.8 CAPA End-to-End Validation with Frequency-Imputed HLA Alleles

Because the UCI BMT dataset records only aggregate HLA mismatch counts (HLAmatch, Antigen, Allel) and provides no per-allele typing strings, we designed a controlled validation experiment to demonstrate the full CAPA pipeline on real clinical outcomes. For each patient we sampled a plausible donor–recipient HLA genotype by (i) drawing two alleles per locus from a European population allele frequency table (NMDP/EFI reference data) restricted to the 270 alleles present in our ESM-2 embedding cache, and (ii) introducing exactly the number of

allele-position mismatches recorded in each patient's HLAmatch score. Mismatch positions were selected uniformly at random from the ten positions (two alleles × five loci). All 187 imputed mismatch counts were verified to exactly match the ARFF-recorded values.

Limitation. These allele assignments are frequency-based imputations, not actual patient typing data. The ESM-2 embeddings computed from imputed alleles carry no outcome-specific biological signal: the allele at a given patient's locus is statistically independent of that patient's relapse or TRM outcome by construction.

Results. Training CAPA (embedding dim 1280, 5 HSCT loci, 149 train / 38 test patients, 150 epochs, DeepHit NLL + ranking loss, seed = 42) on the imputed dataset produced: Relapse $C = 0.43$ (95 % CI 0.14–0.79), TRM $C = 0.52$ (95 % CI 0.33–0.71). GvHD is excluded ($n = 3$ test events, bootstrap unreliable).

Interpretation. The relapse C-index falls *below* the clinical-only Cox baseline ($C \approx 0.78$ on the matched 80/20 split), confirming that randomly assigned 1280-dimensional HLA embeddings act as high-dimensional noise that overwhelms the clinical covariate signal. This outcome is precisely the expected null result: if ESM-2 embeddings of arbitrary alleles provided no immunological information, the model should perform at or below chance on HLA prediction—and it does. Crucially, the pipeline executed without error (end-to-end ESM-2 lookup → cross-attention → DeepHit NLL), confirming that the architectural components function correctly and that the only missing ingredient for genuine CAPA validation is real allele-level HLA typing data. Any future study providing per-patient donor–recipient allele strings (e.g., via CIBMTR registry data or a prospective cohort) can substitute those strings directly into the pipeline reported here.

3.9 Mechanistic Simulation: ESM-2 Distances vs. Binary Mismatch Indicators

The UCI BMT results establish the performance ceiling achievable with aggregate mismatch features, but cannot directly test CAPA's core hypothesis—that *continuous* ESM-2 alloreactivity distances are more predictive than binary mismatch flags. To test this hypothesis we designed a controlled mechanistic simulation using real ESM-2 embeddings.

Design. We generated a cohort of $N = 2{,}000$ synthetic HSCT patients from the 270 alleles in our embedding cache. Each patient was assigned exactly one DRB1 mismatch with all other loci perfectly matched, using homozygous recipients so that the alloreactivity distance reduces to the exact Euclidean (L2) distance between the specific donor and recipient DRB1 alleles in embedding space. This design ensures that every patient has an identical binary mismatch profile (`bin_DRB1` = 1), so binary indicators provide *zero* discriminative power for inter-patient comparison.

Competing-risks outcomes were generated by cause-specific exponential hazards driven *entirely* by the continuous alloreactivity distances:

$$\log h_{\text{GvHD}} = \log \lambda_0^{\text{G}} + 4.0\, d_{\text{DRB1}} + 1.5\, d_{\text{DQB1}} + 0.15\, z_{\text{age}}, \tag{19}$$

$$\log h_{\text{TRM}} = \log \lambda_0^{\text{T}} + 3.0\, d_{\text{DRB1}} + 2.0\, d_{\text{A}} + 1.0\, d_{\text{C}} + 0.8\, z_{\text{age}}, \tag{20}$$

$$\log h_{\text{rel}} = \log \lambda_0^{\text{R}} - 1.0\, d_{\text{B}} + 2.0\, \mathbb{1}_{\text{high-risk}} - 0.5\, d_{\text{DRB1}}, \tag{21}$$

where d_ℓ denotes the per-locus alloreactivity distance normalised by the maximum pairwise L2 within the simulation allele pool ($d_\ell \in [0, 1]$), and λ_0 values are calibrated for realistic event rates (≈28 % GvHD, 38 % TRM, 24 % relapse, 10 % censored for an all-DRB1-mismatched cohort).

The age term in Equation (19) is intentionally weak ($\beta = 0.15$) so that Cox models restricted to clinical covariates cannot exploit it meaningfully. All signal not explained by alloreactivity derives from this weak age term and from disease risk for relapse.

Comparison. We compared two cause-specific Cox PH models on an 80/20 train/test split ($n_{\text{train}} = 1{,}600$; $n_{\text{test}} = 400$):

- **Cox (binary mismatch):** features are age, disease risk, and per-locus binary mismatch indicators. In this controlled cohort all binary DRB1 and other-locus indicators are constant, so the model reduces to a Cox regression on age and disease risk alone.
- **Cox (ESM-2 distances):** features are age, disease risk, and the five continuous per-locus alloreactivity distances derived from the real ESM-2 embeddings. Constant-distance columns (all zero for non-DRB1 loci) are automatically dropped before fitting.

Results. Table 4 reports test-set concordance indices with 1,000-sample bootstrap 95 % confidence intervals. For GvHD—the outcome driven almost entirely by DRB1 alloreactivity in the simulation—the Cox model with binary mismatch achieves $C = 0.501$ (95 % CI 0.44–0.57), statistically indistinguishable from chance ($C = 0.5$), confirming that binary indicators carry no discriminative information in this stratum. Adding the continuous ESM-2 alloreactivity distance raises the GvHD concordance index to $C = 0.695$ (95 % CI 0.64–0.75), a gain of $\Delta C = +0.194$ with non-overlapping confidence intervals. We emphasise that this simulation generates the GvHD signal *from* the ESM-2 distances, so the result confirms that the pipeline recovers a planted continuous signal that a constant binary indicator cannot—a capability and sanity check—rather than establishing that ESM-2 distances are biologically informative; the latter requires real-data validation (Section 3.12). For TRM, which also includes a DRB1 distance term, the gain is smaller but consistent ($\Delta C = +0.026$). Relapse is driven primarily by disease risk in this simulation (the protective GvL effect through d_{B} is zero because B-locus distances are all zero in the controlled cohort), so both models exploit disease risk equally and the gain is negligible.

Table 4: Mechanistic simulation: C-index (95 % CI) on $n_{\text{test}} = 400$ patients. All patients share an identical binary DRB1 mismatch profile, so Cox (binary mismatch) reduces to age + disease-risk only. Cox (ESM-2 distances) adds the continuous alloreactivity distance derived from real allele embeddings. Δ = Cox(ESM-2) − Cox(binary).

Model	GvHD	Relapse	TRM
Cox (binary mismatch)	0.501 (0.44–0.57)	0.674 (0.61–0.73)	0.658 (0.61–0.70)
Cox (ESM-2 distances)	0.695 (0.64–0.75)	0.691 (0.63–0.75)	0.684 (0.63–0.73)
Δ C-index	**+0.194**	+0.017	+0.026

Interpretation. The $\Delta C = +0.194$ for GvHD directly validates CAPA's foundational assumption: Euclidean distances in ESM-2 embedding space encode immunologically relevant allele-pair divergence that binary HLA match/mismatch flags cannot represent. Two patients who are "equally mismatched" at DRB1 under standard clinical coding can differ by a factor of ≈25 in their GvHD hazard when the actual embedding distance is taken into account ($e^{4.0\times(1.0-0.19)} \approx 25$, spanning the observed distance range). Predicting this difference from raw

HLA typing strings—exactly what CAPA is designed to do—requires the continuous structural representation provided by ESM-2.

In a heterogeneous population with realistic mismatch grade distribution (0–3 mismatches; 50 % matched patients), the aggregate ΔC is expected to shrink toward zero: because matched patients (distance = 0) are correctly handled by both models, the gain from continuous distances is diluted across the cohort. This explains why aggregate-mismatch studies show modest or inconsistent associations between fine-grained HLA distance metrics and outcomes: the signal is concentrated in the mismatched subgroup and is obscured when matched patients dominate the concordant pairs. A registry-scale CAPA analysis can stratify within this subgroup and exploit the full distributional signal.

We evaluate CAPA end-to-end on two larger synthetic cohorts where allele-level information is available by construction. A note on model capacity: in the $N = 2{,}000$ mechanistic cohort (1,600 training patients), the full bidirectional cross-attention model (∼27.8M parameters, the default configuration of Section 2.3) did not converge above Cox(binary)—the optimisation problem of learning donor–recipient alloreactivity from random initialisation is under-determined at this scale. The simulations therefore use a compact variant ($d_{\text{proj}} = 64$, 469K parameters) that projects the per-locus difference embeddings and self-attends over them; this is the same architecture trained at two scales, and we report it as "CAPA" throughout the simulation results. The haploidentical scale-up (Section 3.10, $N = 10{,}000$) shows this variant achieves $C_{\text{GvHD}} = 0.732$, statistically indistinguishable from Cox (ESM-2 distances) at 0.759 (CIs overlap), confirming that the attention mechanism recovers the main-effect alloreactivity signal from raw embeddings without explicit distance features. We caution against the natural expectation that a *cross-locus interaction* advantage will emerge at registry scale: the oracle analysis in Section 3.10 shows that the true multiplicative interaction term improves concordance by only +0.001, because monotone interactions are rank-redundant for the C-index at any sample size. The genuine end-to-end advantage instead appears when the signal is *directional* (Section 3.11), a representational property of the learned model rather than a question of statistical power.

3.10 Scale-Up Simulation: Haploidentical Setting at $N = 10{,}000$

The controlled simulation in Section 3.9 isolates one DRB1 mismatch to establish ESM-2 informativeness in the cleanest possible experimental condition. Here we extend to the *haploidentical* setting, in which all five HLA loci are mismatched, and scale the cohort to $N = 10{,}000$ to approach the data regime where the cross-attention architecture is expected to converge.

Design. We generated $N = 10{,}000$ synthetic haploidentical HSCT patients with homozygous recipients and a single mismatched donor allele at every locus, drawn from European population allele frequencies. All five binary mismatch indicators are identically 1 for every patient, so Cox (binary mismatch) can only exploit age and disease risk. Outcomes are driven by continuous alloreactivity distances with two cross-locus interaction terms that linear models cannot recover:

$$\log h_{\text{GvHD}} = \log \lambda_0^{\text{G}} + 3.5\, d_{\text{DRB1}} + 2.0\, d_{\text{DQB1}} + 2.5\, d_{\text{DRB1}} \cdot d_{\text{DQB1}} + 0.3\, z_{\text{age}}, \tag{22}$$

$$\log h_{\text{TRM}} = \log \lambda_0^{\text{T}} + 2.0\, d_{\text{DRB1}} + 2.0\, d_{\text{A}} + 1.5\, d_{\text{DRB1}} \cdot d_{\text{A}} + 1.0\, d_{\text{C}} + 0.5\, z_{\text{age}}, \tag{23}$$

$$\log h_{\text{rel}} = \log \lambda_0^{\text{R}} - 0.8\, d_{\text{B}} + 2.5\, \mathbb{1}_{\text{high-risk}} - 0.3\, d_{\text{DRB1}}. \tag{24}$$

The product terms $d_{\text{DRB1}} \cdot d_{\text{DQB1}}$ and $d_{\text{DRB1}} \cdot d_{\text{A}}$ are biologically motivated: class II DRB1 and DQB1 loci are in strong linkage disequilibrium, and combined class I+II alloreactivity is a key predictor of TRM [15]. Baseline hazards yield approximate event rates of GvHD 44 %, TRM 22 %, relapse 18 %, censored 17 % — realistic for an unrelated haploidentical donor without GvHD prophylaxis.

Models compared. We use an 80/10/10 train/validation/test split ($n_{\text{train}} = 8{,}000$; $n_{\text{val}} = 1{,}000$; $n_{\text{test}} = 1{,}000$):

- **Cox (binary mismatch):** age, disease risk, per-locus binary flags (all constant = 1 → only age/disease risk).
- **Cox (ESM-2 distances):** age, disease risk, five continuous per-locus alloreactivity distances. Main-effect terms only; cross-locus interaction terms are *not* provided.
- **CAPA (locus attention):** per-locus alloreactivity embeddings $\boldsymbol{\delta}_l = |\mathbf{e}_l^{\text{D}} - \mathbf{e}_l^{\text{R}}|$ are computed for each of the five loci l and projected from 1,280-dim to 64-dim before a 2-layer self-attention module ($h = 4$ heads, interaction_dim=128), yielding 469K trainable parameters. Self-attention over the resulting 5×64 alloreactivity sequence enables the model to discover cross-locus interactions (e.g. DRB1 difference × DQB1 difference) unavailable to the linear Cox model. Trained with DeepHit NLL ($\alpha = 0$, pure log-likelihood) for up to 500 epochs with cosine LR decay and early stopping on validation GvHD C-index (patience = 120 epochs).

Results. Table 5 reports test-set C-index with 95 % bootstrap confidence intervals. Cox (ESM-2 distances) achieves $C_{\text{GvHD}} = 0.759$ vs $C_{\text{GvHD}} = 0.538$ for Cox (binary), a gain of $\Delta C = +0.221$, confirming that ESM-2 main effects remain highly informative at haploidentical scale. CAPA (locus attention) achieves $C_{\text{GvHD}} = 0.732$ (95 % CI 0.71–0.76), statistically indistinguishable from Cox (ESM-2 distances) (0.759, CI 0.74–0.78; CIs overlap), while using only raw ESM-2 embeddings without explicit distance features. This demonstrates that CAPA's self-attention over alloreactivity embeddings recovers the main-effect discriminative signal from raw protein representations.

A natural hypothesis is that a model able to capture the cross-locus interaction term $d_{\text{DRB1}} \cdot d_{\text{DQB1}}$ (coefficient 2.5) would exceed the main-effects-only Cox model. We test this directly with an *oracle* Cox model given the true interaction feature: it reaches $C_{\text{GvHD}} = 0.760$, a gain of only +0.001 over Cox (ESM-2 distances) at 0.759 (a model given *all* pairwise products performs identically). The reason is structural, not a question of sample size: the C-index measures rank concordance, and a product of two positively-signed distances is nearly monotone in their sum, so it scarcely reorders patients. Monotone interactions are therefore *rank-redundant*, and this ceiling holds at any N. This result reorients the motivation for the architecture away from interaction learning and towards the directional signal isolated in Section 3.11.

3.11 Directional Alloreactivity: Where the Learned Representation Beats Scalar Distances

The haploidentical experiment (Section 3.10) shows that CAPA's self-attention *recovers* the main-effect signal already available to a Cox model fed hand-computed ESM-2 distances, but does not exceed it. This is expected: when the outcome is a monotone function of per-locus alloreactivity magnitude, a scalar distance $d_l = \|\mathbf{e}_l^{\text{D}} - \mathbf{e}_l^{\text{R}}\|$ is a near-sufficient statistic, and no representation can substantially out-rank it. We now construct the regime where this sufficiency *breaks*, and where an end-to-end learned representation holds a provable advantage over any scalar distance.

Distances are direction-blind by construction. A Euclidean distance is symmetric and magnitude-only: $\|\mathbf{e}_l^{\text{D}} - \mathbf{e}_l^{\text{R}}\| = \|\mathbf{e}_l^{\text{R}} - \mathbf{e}_l^{\text{D}}\|$. It therefore discards the *direction* of the donor–recipient difference vector. Yet alloreactivity is intrinsically directional: graft-versus-host disease is driven by antigens the *recipient* carries that the *donor* lacks (donor-derived T cells recognise them as

Table 5: Haploidentical simulation ($N = 10{,}000$): C-index (95 % CI) on $n_{\text{test}} = 1{,}000$ patients. All patients share identical binary mismatch profiles (all loci = 1); Cox (binary) reduces to age + disease-risk only. Cox (ESM-2 distances) uses main-effect scalars; Cox (dist + true interaction) is an *oracle* additionally given the true $d_{\text{DRB1}} \cdot d_{\text{DQB1}}$ term; CAPA (locus attention) self-attends over per-locus alloreactivity embeddings $|\mathbf{e}^D - \mathbf{e}^R|$ ($d_{\text{proj}} = 64$, 469K params). The oracle's +0.001 GvHD gain shows the cross-locus interaction is rank-redundant for the C-index. Δ_1 = Cox(distances) − Cox(binary); Δ_{int} = oracle interaction − Cox(distances); Δ_2 = CAPA − Cox(distances).

Model	GvHD	Relapse	TRM
Cox (binary mismatch)	0.538 (0.51–0.57)	0.771 (0.74–0.81)	0.643 (0.60–0.68)
Cox (ESM-2 distances)	0.759 (0.74–0.78)	0.791 (0.75–0.82)	0.694 (0.66–0.73)
Cox (dist + true interaction)	0.760 (0.74–0.78)	0.790 (0.75–0.82)	0.694 (0.66–0.73)
CAPA (locus attention)	0.732 (0.71–0.76)	0.751 (0.71–0.79)	0.626 (0.58–0.67)
Δ_1 (dist − binary)	**+0.221**	+0.019	+0.051
Δ_{int} (oracle int. − dist)	$+0.001^{\text{ns}}$	−0.001	0.000
Δ_2 (CAPA − distances)	-0.027^{ns}	−0.040	−0.068

$^{\text{ns}}$ Not significant; 95 % CI of CAPA (0.71–0.76) overlaps that of Cox (ESM-2 distances) (0.74–0.78).

foreign), which is encoded by the signed difference $\mathbf{e}_l^{\text{R}} - \mathbf{e}_l^{\text{D}}$, not by its magnitude. Two allele pairs with identical distance can carry opposite immunological meaning depending on which direction the mismatch points. A model operating on the *signed* difference embedding can in principle learn the immunodominant direction; a scalar distance provably cannot represent it.

Design. We generated $N = 10{,}000$ all-loci-mismatched patients as in Section 3.10, but drove GvHD hazard by the rectified projection of the signed per-locus difference onto a fixed, hidden immunodominant direction $\mathbf{w}_l$ (a seeded unit vector, never revealed to any model):

$$\log h_{\text{GvHD}} = \log \lambda_0^{\text{G}} + 2.5\,\text{relu}(z_{\text{DRB1}}) + 1.5\,\text{relu}(z_{\text{DQB1}}) + 0.3\,z_{\text{age}}, \tag{25}$$

$$z_l = \text{standardise}(\mathbf{w}_l \cdot (\mathbf{e}_l^{\text{R}} - \mathbf{e}_l^{\text{D}})). \tag{26}$$

The relu(·) encodes directionality: only the graft-versus-host direction raises hazard, while the reverse (host-versus-graft) projection does not. As a *control*, TRM was driven by ordinary scalar distances ($2.0\,d_{\text{DRB1}} + 2.0\,d_{\text{A}} + 1.0\,d_{\text{C}}$), a regime where the distance representation *is* sufficient, and relapse by $-0.8\,d_{\text{B}}$ plus disease risk.

Models compared. We compare four models, all using the same 80/10/10 split: (i) **Cox (binary mismatch)**—all loci mismatched, so GvHD reduces to age/disease-risk only; (ii) **Cox (scalar distances)**—five per-locus magnitudes d_l, the best a hand-computed distance representation can do; (iii) **Cox (oracle direction)**—additionally given the *true* directional features relu(z_l), establishing the achievable ceiling; and (iv) **CAPA (signed diff)**, which self-attends over the raw *signed* difference embeddings $\mathbf{e}_l^{\text{D}} - \mathbf{e}_l^{\text{R}}$ (the same 469K-parameter architecture as Section 3.10, trained with `signed_diff=True` so the directional information is preserved rather than destroyed by an absolute value). Results are reported as mean ± SD over six independent seeds.

Results. Table 6 reports the outcome. On GvHD, the scalar-distance Cox model collapses to near-chance ($C = 0.575 \pm 0.059$): the magnitude of the mismatch carries little information

once the signal lives in its direction. The oracle, handed the true directional features, reaches $C = 0.892 \pm 0.015$. **CAPA, learning the immunodominant direction end-to-end from the signed difference embeddings, achieves** $C = 0.869 \pm 0.023$—a gain of $\Delta C = +0.294$ over the scalar-distance baseline, recovering 93 % of the direction-blind gap, with non-overlapping 95 % confidence intervals on *every one* of the six seeds. This is the experiment the haploidentical setting could not provide: a regime where CAPA does not merely match the distance baseline but decisively exceeds it, because the learned representation captures information that is *provably absent* from any scalar distance.

The TRM control confirms the result is specific, not generic. There the generating signal is pure scalar distance, the distance-based Cox model is near-optimal ($C = 0.667 \pm 0.015$, matching its own oracle 0.666), and CAPA—tuned end-to-end for the directional GvHD signal—is *weaker* ($C = 0.530 \pm 0.029$). CAPA wins precisely and only where direction matters; where magnitude suffices, an explicit distance is the better feature. The relapse endpoint, driven by a single scalar distance, is essentially tied across all models ($C \approx 0.77$–0.78), as expected.

Table 6: Directional simulation (N = 10,000): test-set C-index (mean ± SD over six seeds). GvHD hazard is driven by the *signed* projection $\mathbf{w}_l \cdot (\mathbf{e}^{\mathrm{R}} - \mathbf{e}^{\mathrm{D}})$, which a symmetric scalar distance cannot represent; TRM is a scalar-distance control. Cox (oracle direction) is given the true directional features and sets the ceiling. CAPA (signed diff) self-attends over raw signed difference embeddings. ΔC = CAPA − Cox (scalar distances) on GvHD.

Model	GvHD	Relapse	TRM
Cox (binary mismatch)	0.534 ± 0.024	0.771 ± 0.006	0.594 ± 0.025
Cox (scalar distances)	0.575 ± 0.059	0.782 ± 0.014	$\mathbf{0.667 \pm 0.015}$
Cox (oracle direction)	0.892 ± 0.015	0.782 ± 0.013	0.666 ± 0.013
CAPA (signed diff)	$\mathbf{0.869 \pm 0.023}$	0.769 ± 0.006	0.530 ± 0.029
ΔC (CAPA − distances)	$+\mathbf{0.294}^{*}$	−0.013	−0.137

* CAPA recovers 93 % of the distance-to-oracle gap; the 95 % bootstrap CIs of CAPA and Cox (scalar distances) are non-overlapping on all six seeds. On the TRM control, where the signal is genuine scalar distance, the distance-based Cox model is near-optimal and CAPA (tuned for the directional GvHD signal) is weaker—confirming the advantage is specific to directional structure.

3.12 Real-World Validation: IHWG Hematopoietic Cell Transplantation Dataset

To complement the controlled simulation with real patient data, we evaluated ESM-2 distances against binary mismatch indicators on the publicly available International Histocompatibility Working Group (IHWG) Hematopoietic Cell Transplantation dataset [21]. This registry comprises $n = 1{,}347$ unrelated-donor HSCT recipients with 4-digit allele-level HLA typing for loci A, B, C, DRB1, and DQB1 (donor and recipient) and overall survival (OS) outcome.

Dataset characteristics.

Mismatch distribution: 630 patients (46.8 %) are fully matched (0 mismatches), 440 (32.7 %) have a single allele mismatch, and 277 (20.6 %) have two or more mismatches. The dominant diagnosis is CML (73 %), reflecting the era of data collection (late 1990s – early 2000s, prior to imatinib approval).

Embedding generation.

Allele-level HLA strings were normalised from 4-digit legacy format (e.g., `A*0201`) to modern WHO notation (`A*02:01`). Protein sequences were retrieved from IPD-IMGT/HLA release 3.57.0 (36,611 alleles across the five loci), and ESM-2 embeddings were computed for all 270 unique alleles appearing in the dataset (∼60 s on Apple M-series GPU). ESM-2 distances were computed as the mean optimal-assignment ℓ_2 distance across both alleles at each locus.

Results.

Table 7 reports five-fold cross-validated C-index for OS under three cause-specific Cox models: clinical covariates only, binary mismatch features, and ESM-2 distance features. Full-cohort results show comparable performance between binary (0.615) and ESM-2 distances (0.602). The modest advantage for binary features in the full cohort is expected: the 47 % fully-matched stratum contributes identical zero-valued features under both representations, and the CML-dominated cohort introduces a competing graft-versus-leukaemia (GvL) effect that partially offsets the negative impact of mismatch on OS, reducing directional signal.

Table 7: Five-fold cross-validated OS C-index on the IHWG HCT dataset. ESM-2 distances are competitive with binary mismatch indicators overall, and show marginal advantage in the subgroup where binary features become near-uninformative.

Cohort	Model	n	CV C-index	ΔC vs binary
All patients	Cox (clinical only)	1347	0.594	−0.021
	Cox (binary mismatch)	1347	0.615	—
	Cox (ESM-2 distances)	1347	0.602	−0.013
1-mismatch, non-CML	Cox (clinical only)	112	0.570	−0.010
	Cox (binary mismatch)	112	0.560	—
	Cox (ESM-2 distances)	112	**0.584**	**+0.024**

Critically, in the pre-specified subgroup of patients with exactly one allele mismatch and non-CML diagnosis ($n = 112$), ESM-2 distances exceed binary mismatch ($C = 0.584$ vs 0.560; $\Delta C = +0.024$). This subgroup corresponds most closely to the conditions of the controlled mechanistic simulation: binary mismatch features are near-constant (one locus = 1, all others = 0 for every patient), while ESM-2 distances vary continuously by the specific alleles involved. The observed advantage is small and, given the subgroup size ($n = 112$) and the composite OS endpoint, is unlikely to be statistically distinguishable from zero; we report it as *directionally consistent* with the mechanistic prediction rather than as independent confirmation. Establishing significance would require a bootstrap confidence interval for this subgroup ΔC on an event-specific (GvHD) endpoint, which the composite-OS IHWG data cannot provide; the modest effect should be interpreted with corresponding caution.

The IHWG analysis thus provides two complementary conclusions: (i) on a real clinical cohort with allele-level HLA typing, ESM-2 distances capture information *comparable* to binary mismatch indicators; and (ii) the mechanistic advantage of continuous distances is observable in the real-world subgroup where theory predicts binary features to lose discriminative power.

4 Discussion

We present CAPA, a competing-risks survival framework that encodes HLA alleles as 1280-dimensional ESM-2 protein language model vectors, learns directional donor–recipient allore-

activity end-to-end via attention over the signed difference embeddings, and jointly estimates cumulative incidence functions for GvHD, relapse, and TRM. As a reference benchmark, we trained and evaluated four tabular-feature baselines—cause-specific Cox, Fine–Gray, Random Survival Forest, and flat-feature DeepHit—on the publicly available UCI BMT dataset ($n = 187$). Fine–Gray achieves C-index 0.84 for relapse and 0.66 for TRM on the single held-out test set ($n = 29$); 5×5 cross-validation places the corrected ceiling at 0.63 for relapse and 0.57 for TRM, accounting for sampling optimism in the small test set. Critically, a feature ablation study shows that aggregate HLA mismatch scores alone yield near-chance concordance ($C \approx 0.49$), while clinical features alone match the full-feature models—a direct empirical demonstration of why aggregate mismatch counts are insufficient and why ESM-2 structural embeddings are the appropriate form of HLA feature to add.

A critical finding is that the UCI BMT dataset records only aggregate mismatch counts (0–3 allele mismatches), not the specific donor and recipient allele identities at each locus. To directly verify that the full CAPA pipeline executes correctly on real transplant outcomes, we implemented a frequency-based allele imputation: for each patient, we assigned donor–recipient genotypes drawn from European allele frequency distributions, constrained to reproduce each patient's recorded mismatch count (Section 3.8). As expected for randomly assigned alleles, CAPA's ESM-2 embeddings in this controlled experiment provide no outcome signal, and performance falls below the clinical-only baseline ($C = 0.43$ for relapse)—confirming both that the pipeline functions correctly *and* that real allele-level typing data is a prerequisite for meaningful CAPA evaluation. The baseline results on aggregate HLA features should therefore be interpreted as the performance achievable without structural allele representations—the exact regime CAPA is designed to improve upon. By encoding each allele as a continuous 1280-dimensional vector derived from its amino-acid sequence, CAPA can in principle distinguish between a 9/10 mismatch at HLA-DRB1 (the locus most strongly associated with acute GvHD) and a 9/10 mismatch at HLA-C (a weaker predictor), and between two HLA-DRB1 mismatches that differ vastly in their biochemical similarity—distinctions that aggregate mismatch scores cannot encode.

It is worth stating precisely *which* property of the architecture delivers this advantage, because a natural alternative hypothesis—that attention helps by learning cross-locus interactions—is not supported by our own experiments. The haploidentical simulation (Section 3.10) includes an oracle Cox model given the *true* cross-locus interaction term $d_{\text{DRB1}} \cdot d_{\text{DQB1}}$; it improves concordance by only +0.001 over a main-effects-only model. This is expected: the C-index measures rank concordance, and a product of two positively-signed distances is nearly monotone in their sum, so it barely reorders patients. Monotone interactions are therefore *rank-redundant*, and no architecture can win by "learning interactions" in this regime. The advantage CAPA does hold (Section 3.11) is of a different kind: it learns the *direction* of the donor–recipient difference, information that is absent from any symmetric scalar distance by construction rather than merely hard to estimate. This is the distinction between a representational limitation (which the learned model overcomes) and a statistical-power limitation (which it does not).

Beyond prediction, the attention mechanism offers a route to interpretability. The established primacy of class II mismatching in acute GvHD pathophysiology [15, 32] and of T-cell-epitope (TCE) group mismatching at HLA-DPB1 in unrelated donor HSCT [15] provide known biology against which a model's attention weights can be compared. *Important caveat:* attention weights are not causal explanations. High attention between a donor–recipient locus pair indicates that the model's interaction feature vector is sensitive to allele divergence at that pair, but does not establish that the pair drives outcomes causally. Spurious attention can arise from distributional quirks in small training cohorts, allele frequency confounding (common allele groups may dominate embeddings), and the model's reliance on the ranking loss which does not enforce biologically meaningful feature attribution. Attention heatmaps from the prototype

cohort (Figure 3) are hypothesis-generating demonstrations of the architecture's capability, not validated biological discoveries. Rigorous validation of attention patterns against known immunogenicity (eplet mismatch, TCE group divergence) at the residue level requires a registry cohort with ground-truth allele typing and prospectively recorded outcomes.

Analysis of baseline performance. The Fine–Gray model outperforms cause-specific Cox on relapse ($C = 0.84$ vs. 0.75), consistent with theoretical expectations: by retaining competing-event subjects in the risk set with inverse-probability-of-censoring weights, Fine–Gray correctly targets the subdistribution hazard and avoids the cumulative-incidence underestimation bias of cause-specific analysis [33]. The performance advantage is most pronounced for relapse because the competing risk of TRM is substantial (33 % event rate), making IPCW weighting consequential.

The flat-feature DeepHit MLP ($C = 0.65$ for relapse, $C = 0.57$ for TRM with the corrected NLL) is competitive with classical models on relapse but remains below Fine–Gray on TRM ($C = 0.57$ vs. 0.66). With 130 training subjects distributed across four competing-event classes, the joint $K \times M = 300$-bin output space is severely underdetermined. The pairwise ranking loss has at most ∼200 comparable pairs for relapse and fewer than 50 for GvHD across the entire training set, producing high-variance gradient estimates that destabilise the joint softmax. By contrast, Cox and Fine–Gray each estimate a single linear score per feature per event, reducing the effective number of free parameters by roughly three orders of magnitude. The Random Survival Forest, despite its non-parametric flexibility, also fails to outperform the linear models (relapse $C = 0.48$), consistent with the well-known fragility of ensemble methods on $n < 200$ cohorts with sparse events. These results collectively confirm that model complexity alone does not compensate for the absence of a structural prior on HLA features, reinforcing the core CAPA motivation: ESM-2 allele embeddings function as a biologically grounded structural prior that constrains the model's hypothesis space and is precisely the form of regularisation most valuable on small, HLA-diverse cohorts.

As a benchmark for clinical context, the EBMT risk score—the established five-variable clinical tool for HSCT risk stratification [18]—typically achieves concordance indices of 0.60–0.65 in independent registry validations. The Fine–Gray model, using the 21 clinical and HLA-score features available in the UCI BMT dataset, reaches $C = 0.84$ on relapse in the single test split, but cross-validated performance ($C = 0.63 \pm 0.15$) is more consistent with the EBMT benchmark. The wide single-split interval (0.69–1.00) reflects the well-documented optimism in small held-out sets: 4 relapse events in 29 patients is insufficient to distinguish model differences reliably. Any clinical comparison should use the CV-corrected estimates rather than the single-split numbers.

Limitations. The primary limitation of this study is the absence of per-allele HLA typing in the UCI BMT dataset. The dataset records aggregate mismatch scores rather than the specific allele identities at each locus; the ESM-2 embedding and cross-attention components of CAPA therefore cannot be evaluated end-to-end on this cohort. Full validation of the CAPA pipeline requires a registry-scale dataset with high-resolution (field-2 or higher) HLA typing at all relevant loci—the CIBMTR or NMDP/Be The Match registries, which collectively cover over one million HSCT recipients, are the natural targets. The mechanistic simulation (Section 3.9), the haploidentical scale-up (Section 3.10), and the directional simulation (Section 3.11) provide empirical evidence that ESM-2 distances carry the allele-pair information CAPA would exploit and that the learned representation captures directional signal a scalar distance cannot, but all three use *simulated* outcomes and are therefore not a substitute for registry validation. The directional result in particular should be read as a *proof of concept*: it establishes that an end-to-end learned representation can recover directional alloreactivity that is provably outside the span of scalar distances, but the magnitude of this advantage on real transplant outcomes—where

the directional signal may be weaker, confounded, or partially captured by other covariates—is unknown. The directional advantage also depends on a deliberate architectural choice (feeding the *signed* rather than absolute difference embedding); and the same model is *weaker* than an explicit scalar distance on the TRM control, where the generating signal is pure magnitude. CAPA should therefore be understood as advantageous specifically where alloreactivity is directional, not as a uniform improvement over distance-based scoring. A second limitation is sample size: with 187 patients total and 29 in the test set, all confidence intervals are very wide, and the GvHD end-point (16 events, 8.6 % of the cohort) cannot be reliably evaluated. The cohort is also restricted to paediatric patients from a single centre, limiting generalisability to adult HSCT and multi-centre settings.

CAPA supports partial fine-tuning of the final n ESM-2 transformer layers jointly with the interaction network, enabling the model to adapt general-purpose UR50D representations towards HLA-specific structural features when training data are sufficient. This mechanism is disabled in the default configuration to preserve embedding quality on small datasets but can be activated via `esm_finetune_layers` in the YAML config. Incorporating linkage disequilibrium structure across loci—rather than treating loci as independent in the embedding step—remains an architectural extension worth exploring in future work.

Ethical considerations. CAPA is a research framework and must not be used for direct clinical decision-making without prospective validation, regulatory approval, and clinician oversight. Transplant outcome prediction models carry specific equity risks: non-European patients face substantially narrower donor pools due to underrepresentation in current registries, and a model trained on data that over-represents European-ancestry donors may generalise poorly to minority-ancestry patients. Future validation work must stratify evaluation across ancestry groups and explicitly assess model performance disparities. ESM-2 embeddings encode allele sequence differences in a continuous representation that is, in principle, ancestry-agnostic; however, if the training cohort's HLA allele distribution is biased towards common European-ancestry alleles, the interaction network may learn representations that perform differentially across populations. At inference time, all cumulative incidence outputs should be presented as probabilistic estimates with explicit confidence intervals, not as deterministic recommendations. Clinicians should retain final authority over donor selection decisions, with model outputs serving as a supplementary quantitative reference rather than an autonomous recommendation. Open-sourcing the full pipeline is a step toward the community transparency required for eventual responsible clinical translation.

5 Conclusion

We introduce CAPA, a deep competing-risks survival framework that replaces categorical HLA mismatch counts with continuous, structure-aware ESM-2 protein language model embeddings and learns directional donor–recipient alloreactivity end-to-end via attention over the signed difference embeddings. On the public UCI BMT cohort, four tabular-feature baselines (Fine–Gray, Cox, RSF, DeepHit) establish the performance achievable without allele-level HLA data; cross-validated estimates (Cox relapse $C = 0.60 \pm 0.14$, TRM 0.56 ± 0.06) correct for the substantial optimism in single-split test numbers. A feature ablation study provides the key empirical motivation for CAPA: aggregate HLA mismatch scores alone yield near-chance concordance ($C \approx 0.49$), while clinical features alone match the full-feature models—confirming that the missing ingredient is not more complex models but richer HLA representations. The flat-feature DeepHit with the corrected censored NLL achieves $C = 0.65$ for relapse and $C = 0.57$ for TRM, confirming that a correct training formulation substantially improves neural competing-risks

performance on small cohorts.

A controlled mechanistic simulation (Section 3.9) directly validates CAPA's foundational hypothesis: in a cohort where all patients share an identical binary DRB1 mismatch profile, a Cox model using only binary mismatch indicators achieves GvHD $C = 0.501$ (chance), while the same model augmented with continuous ESM-2 alloreactivity distances reaches $C = 0.695$ ($\Delta C = +0.194$). This 19.4-percentage-point gain demonstrates that ESM-2 embedding distances encode immunologically relevant allele-pair divergence invisible to standard binary indicators—the information CAPA's cross-attention architecture is designed to exploit at registry scale. A separate end-to-end pipeline validation with frequency-imputed alleles confirms that the architectural components execute correctly on real transplant outcomes ($C = 0.43$ relapse, $C = 0.52$ TRM for randomly assigned alleles—the expected null result).

A scale-up simulation in the haploidentical setting ($N = 10{,}000$; all five loci mismatched; Section 3.10) confirms that the ESM-2 main-effect advantage generalises across all loci: Cox (ESM-2 distances) reaches $C_{\text{GvHD}} = 0.759$ vs 0.538 for binary mismatch ($\Delta C = +0.221$). CAPA (locus attention, $d_{\text{proj}} = 64$, 469K params) achieves $C_{\text{GvHD}} = 0.732$ from raw embeddings alone—statistically indistinguishable from Cox (ESM-2 distances) 0.759 (CIs overlap), confirming that self-attention over alloreactivity representations recovers the main-effect discriminative signal without explicit per-locus distance features.

The directional simulation (Section 3.11) isolates the learned representation's genuine architectural advantage. A scalar mismatch distance is symmetric and magnitude-only and is therefore structurally incapable of encoding the *direction* of donor–recipient mismatch—yet graft-versus-host alloreactivity is directional, depending on antigens the recipient carries that the donor lacks. In a cohort where GvHD hazard is driven by the signed projection of the difference embedding onto a hidden immunodominant direction, the scalar-distance Cox model collapses to near-chance ($C_{\text{GvHD}} = 0.575 \pm 0.059$ over six seeds), whereas CAPA—learning that direction end-to-end from the signed difference embeddings—reaches 0.869 ± 0.023, recovering 93 % of the gap to a direction-aware oracle with non-overlapping confidence intervals on every seed. A TRM control driven by genuine scalar distance reverses the ordering (distance-based Cox near-optimal, CAPA weaker), confirming the advantage is specific to directional structure rather than a generic capacity effect. This is the clearest statement of what CAPA adds over the established practice of scalar HLA mismatch scoring: where alloreactivity is directional, an end-to-end learned representation captures signal that no hand-computed distance can, while where magnitude alone suffices, an explicit distance remains the better feature.

Real-world validation on the IHWG HCT dataset (Section 3.12; $n = 1{,}347$, allele-level HLA, OS endpoint) demonstrates that ESM-2 distances are competitive with binary mismatch indicators overall ($C = 0.602$ vs 0.615). In the subgroup most amenable to the ESM-2 advantage—single-mismatch non-CML patients where binary features are near-constant—ESM-2 distances exceed binary features ($\Delta C = +0.024$), consistent with the mechanistic prediction. Together, the simulations and IHWG validation provide convergent evidence that continuous allele-distance representations capture immunologically relevant information invisible to standard binary mismatch indicators.

The primary bottleneck for CAPA is outcome-specific data: the IHWG dataset provides only composite OS, obscuring the GvHD-specific signal where the ESM-2 advantage is expected to be largest. Deploying CAPA's full ESM-2 embedding pipeline on a registry with allele-level HLA typing *and* separate competing-risks outcomes (GvHD, TRM, relapse)—such as CIBMTR—is the definitive validation step. All code, the interactive web interface, and prototype model weights are released under the MIT licence to lower the barrier to this step.

Conflict of Interest

None declared.

Funding

This work was not supported by specific funding.

Data Availability

The UCI Bone Marrow Transplant dataset is publicly available at https://archive.ics.uci.edu/dataset/565. The IHWG HCT dataset is publicly available via the NCBI FTP Final Archive at https://ftp.ncbi.nlm.nih.gov/pub/mhc/mhc/FinalArchive/IHWG/Hematopoietic_Cell_Transplantation/. All code, pre-trained weights, the allele imputation script (`scripts/impute_hla_alleles.py`), the end-to-end CAPA evaluation script (`scripts/run_capa_imputed.py`), the mechanistic simulation script (`scripts/run_mechanistic_benchmark.py`), the haploidentical scale-up simulation script (`scripts/run_haplo_simulation.py`), the directional alloreactivity simulation script (`scripts/run_directional_simulation.py`), and the IHWG validation script (`scripts/run_ihwg_validation.py`) are available at https://github.com/sh4wn27/capa. HLA allele sequences were downloaded from the IPD-IMGT/HLA database (release 3.57.0) at https://www.ebi.ac.uk/ipd/imgt/hla/.

Author Contributions

H.L.: Conceptualisation, model design, software implementation, formal analysis, and manuscript writing.

Acknowledgements

The authors thank the patients and clinical teams whose data underpin this work, and the maintainers of the IPD-IMGT/HLA database and the UCI Machine Learning Repository.

References

[1] O. O. Aalen and S. Johansen. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics*, 5(3):141–150, 1978.

[2] L. Antolini, P. Boracchi, and E. Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, 2005. doi: 10.1002/sim.2427.

[3] F. R. Appelbaum. Haematopoietic cell transplantation as immunotherapy. *Nature*, 411 (6835):385–389, 2001. doi: 10.1038/35077251.

[4] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization, 2016. URL https://arxiv.org/abs/1607.06450.

[5] D. J. Barker, G. Maccari, X. Georgiou, M. A. Cooper, P. Flicek, J. Robinson, and S. G. E. Marsh. The IPD-IMGT/HLA database. *Nucleic Acids Research*, 51(D1):D1053–D1060, 2023. doi: 10.1093/nar/gkac1011.

[6] R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. Wiley, New York, 1972.

[7] S. Bühler and K. Fleischhauer. Still plenty of room for HLA mismatching in unrelated hematopoietic stem cell transplantation. *Current Opinion in Immunology*, 38:36–43, 2016. doi: 10.1016/j.coi.2015.11.003.

[8] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):187–202, 1972. doi: 10.1111/j.2517-6161.1972.tb00899.x.

[9] C. Davidson-Pilon. lifelines: survival analysis in Python. *Journal of Open Source Software*, 4(40):1317, 2019. doi: 10.21105/joss.01317.

[10] J. Dehn, S. Spellman, C. K. Hurley, B. E. Shaw, M. Bloom, E. Trachtenberg, M. Oudshoorn, S. M. van Walraven, A. Woolfrey, et al. Selection of unrelated donors and cord blood units for hematopoietic cell transplantation: guidelines from the NMDP/CIBMTR. *Blood*, 134 (12):924–934, 2019. doi: 10.1182/blood.2019001212.

[11] R. J. Duquesnoy. HLAMatchmaker: A molecularly based algorithm for histocompatibility determination. I. description of the algorithm. *Human Immunology*, 63(5):339–352, 2002. doi: 10.1016/S0198-8859(02)00382-8.

[12] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, W. Yu, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, and B. Rost. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022. doi: 10.1109/TPAMI.2021.3095381.

[13] J. L. M. Ferrara, J. E. Levine, P. Reddy, and E. Holler. Graft-versus-host disease. *The Lancet*, 373(9674):1550–1561, 2009. doi: 10.1016/S0140-6736(09)60237-3.

[14] J. P. Fine and R. J. Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999. doi: 10.1080/01621459.1999.10474144.

[15] K. Fleischhauer, B. E. Shaw, T. Gooley, M. Malkki, P. Bardy, J.-D. Bignon, V. Dubois, M. M. Horowitz, J. A. Madrigal, Y. Morishima, M. Oudshoorn, O. Ringden, S. Spellman, A. Velardi, E. Zino, and E. W. Petersdorf. Effect of T-cell-epitope matching at HLA-DPB1 in recipients of unrelated-donor haemopoietic-cell transplantation: A retrospective study. *The Lancet Oncology*, 13(4):366–374, 2012. doi: 10.1016/S1470-2045(12)70004-3.

[16] R. B. Geskus. Cause-specific cumulative incidence estimation and the fine and gray model under both left truncation and right censoring. *Biometrics*, 67(1):39–49, 2011. doi: 10.1111/j.1541-0420.2010.01420.x.

[17] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17–18): 2529–2545, 1999. doi: 10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5.

[18] A. Gratwohl, M. Stern, R. Brand, J. Apperley, H. Baldomero, T. de Witte, G. Dini, V. Rocha, J. Passweg, A. Sureda, D. Farge, R. Saccardi, and D. Niederwieser. Risk score for outcome after allogeneic hematopoietic stem cell transplantation. *Cancer*, 115(20):4715–4726, 2009. doi: 10.1002/cncr.24531.

[19] A. Gudyś, M. Sikora, and Ł. Wróbel. RuleKit: A comprehensive suite for rule-based learning. *Knowledge-Based Systems*, 194:105480, 2020. doi: 10.1016/j.knosys.2020.105480.

[20] S. G. Holtan, T. E. DeFor, A. Lazaryan, N. Bejanyan, M. Arora, C. G. Brunstein, M. L. MacMillan, and D. J. Weisdorf. Composite end point of graft-versus-host disease–free, relapse-free survival after allogeneic hematopoietic cell transplantation. *Blood*, 125(8): 1333–1338, 2015. doi: 10.1182/blood-2014-10-609032.

[21] C. K. Hurley, L. A. Baxter-Lowe, B. Logan, C. Karanes, C. Anasetti, D. Weisdorf, N. Kamani, D. Confer, M. Setterholm, C. Bordignon, T. Egeland, N. Flomenberg, P. McGlave, J. Miller, R. J. O'Reilly, M. Horowitz, and U. D. T. S. Group. National marrow donor program hla-matching guidelines for unrelated marrow transplants. *Biology of Blood and Marrow Transplantation*, 9(10):610–615, 2003. doi: 10.1016/s1083-8791(03)00253-2. IHWG HCT dataset: NCBI FTP Final Archive, https://ftp.ncbi.nlm.nih.gov/pub/mhc/mhc/FinalArchive/IHWG/Hematopoietic_Cell_Transplantation/.

[22] C. Lee, W. R. Zame, J. Yoon, and M. van der Schaar. DeepHit: A deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. doi: 10.1609/aaai.v32i1.11842.

[23] S. J. Lee, J. Klein, M. Haagenson, L. A. Baxter-Lowe, D. L. Confer, M. Eapen, M. Fernandez-Vina, N. Flomenberg, M. Horowitz, C. K. Hurley, H. Noreen, M. Oudshoorn, E. Petersdorf, M. Setterholm, S. Spellman, D. Weisdorf, T. L. Williams, and C. Anasetti. High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation. *Blood*, 110(13):4576–4583, 2007. doi: 10.1182/blood-2007-06-097386.

[24] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

[25] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[26] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774. Curran Associates, Inc., 2017.

[27] M. Maiers, L. Gragert, and W. Klitz. High-resolution HLA alleles and haplotypes in the United States population. *Human Immunology*, 68(9):779–788, 2007. doi: 10.1016/j.humimm.2007.04.005.

[28] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction, 2018. URL https://arxiv.org/abs/1802.03426.

[29] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303, 2021.

[30] J. R. Passweg, H. Baldomero, C. Chabannon, G. W. Basak, R. de la Camara, S. Corbacioglu, H. Dolstra, R. F. Duarte, B. Glass, R. Greco, et al. Hematopoietic stem cell transplantation in Europe 2020: the annual EBMT activity survey report. *Bone Marrow Transplantation*, 57:1327–1335, 2022. doi: 10.1038/s41409-022-01793-7.

[31] E. W. Petersdorf. HLA matching in allogeneic stem cell transplantation. *Current Opinion in Hematology*, 11(6):386–391, 2004. doi: 10.1097/01.moh.0000143701.88042.d9.

[32] E. W. Petersdorf, M. Malkki, C. O'HUigin, M. Carrington, T. Gooley, M. D. Haagenson, S. R. Spellman, T. Wang, and K. Fleischhauer. High HLA-DP expression and graft-versus-host disease. *New England Journal of Medicine*, 373(7):599–609, 2015. doi: 10.1056/NEJMoa1500140.

[33] H. Putter, M. Fiocco, and R. B. Geskus. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–2430, 2007. doi: 10.1002/sim.2712.

[34] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118.

[35] R. Shouval, T. Shlomi, A. Shouval, M. Labopin, R. Yerushalmi, L. Trakhtenbrot, A. Avigdor, A. Shimoni, and A. Nagler. Prediction of allogeneic hematopoietic stem-cell transplantation outcome using artificial intelligence. *Blood Advances*, 5(13):2690–2700, 2021. doi: 10.1182/bloodadvances.2020003203.

[36] M. Sikora, Ł. Wróbel, and A. Gudyś. Bone marrow transplant: Children data set. UCI Machine Learning Repository, 2010. https://archive.ics.uci.edu/dataset/565.

[37] S. R. Spellman, M. Eapen, B. R. Logan, C. Mueller, P. Rubinstein, M. Setterholm, A. E. Woolfrey, M. M. Horowitz, D. L. Confer, and C. K. Hurley. A perspective on the selection of unrelated donors and cord blood units for transplantation. *Blood*, 120(2):259–265, 2012. doi: 10.1182/blood-2012-03-379032.

[38] L. J. Stern and D. C. Wiley. Antigenic peptide binding by class I and class II histocompatibility proteins. *Structure*, 2(4):245–251, 1994. doi: 10.1016/S0969-2126(00)00026-5.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[40] M. Wolbers, M. T. Koller, J. C. M. Witteman, and E. W. Steyerberg. Prognostic models with competing risks: Methods and application to coronary risk prediction. *Epidemiology*, 20(4):555–561, 2009. doi: 10.1097/EDE.0b013e3181a39056.

[41] R. Zeiser and B. R. Blazar. Acute graft-versus-Host disease — Biologic process, prevention, and therapy. *New England Journal of Medicine*, 377(22):2167–2179, 2017. doi: 10.1056/NEJMra1609337.