**Inferring the T-cells repertoire dynamics of healthy individuals**

**Tool/Model Summary Template**

1. **Tool/Model Name:**  
   **Stochastic Clonal Dynamics Model** (Geometric Brownian Motion-based inference model)
2. **Reference(s):**  
   Bensouda Koraichi, M., Ferri, S., Walczak, A. M., & Mora, T. (2022). *Inferring the T-cells repertoire dynamics of healthy individuals*. bioRxiv. https://doi.org/10.1101/2022.05.01.490247
3. **Category:**

* Full Repertoire Simulator
* Clone Dynamics Model
* Statistical Inference from RepSeq
* Longitudinal TCR Analysis

1. **Main Purpose/Function:**

* To infer the turnover dynamics of T-cell receptor (TCR) clones in healthy individuals over time, in the absence of strong antigenic stimulation.
* It estimates clone-specific dynamics using longitudinal immune repertoire sequencing (RepSeq) data and a stochastic modeling framework.

1. **Key Features:**

Probabilistic, Bayesian inference framework

* Models clonal dynamics using **Geometric Brownian Motion (GBM)**
* Corrects for **experimental and biological noise** using replicate data
* Capable of **frequency-dependent analysis**
* Repertoire-wide and clone-level inference
* Applicable without replicates for large clones

1. **Input Requirements:**

* Longitudinal TCR repertoire sequencing data (same individuals, multiple timepoints)
* Clone frequencies (read counts or normalized)
* (Optionally) Replicate samples at same timepoint to estimate noise
* Clone size thresholds (frequency cutoffs)

1. **Output:**

* Inferred parameters:
  + - **τ (tau):** decay/turnover time of clones
    - **θ (theta):** amplitude of stochastic fluctuations
* Posterior distributions for fold-changes in clone sizes
* Predicted clone persistence probabilities
* Exponent α of the clone-size distribution

Visualizations: clone frequency distributions, persistence curves, inferred vs. true parameters (in synthetic tests)

1. **Validation/Evaluation:**

* Validated on **synthetic data** simulating realistic repertoire dynamics and sequencing noise
* Comparison of **naive vs. full inference** methods
* Applied to real datasets from 9 individuals
* Good agreement with empirical clone-size distributions (α ≈ 1) and known aging effects

1. **Strengths:**

* Captures biologically meaningful parameters (τ, θ)
* Can work without replicates for large clones
* Provides insights into age-dependent dynamics
* Compatible with existing RepSeq datasets
* Theoretical grounding in stochastic processes
* Model-predicted persistence validated with data

1. **Limitations:**

* Requires replicate data for small clone accuracy
* Assumes constant dynamics per clone (but some frequency-dependence exists)
* Does not distinguish naive vs. memory phenotypes explicitly
* Assumes consistent thymic output over short timescales
* Sensitive to data preprocessing (e.g., PCR biases in gDNA data)

1. **Applications in Autoimmunity:**

* Not directly tested in autoimmune conditions
* However, the model is **highly suitable** for studying clone stability, memory maintenance, and repertoire turnover in autoimmune settings
* Could be adapted to detect abnormal clonal dynamics or hyperstability in autoimmune diseases

1. **Notable Citations/Use Cases:**

* This specific model is newly introduced in the manuscript
* Builds on prior work:
  + Desponds et al., PNAS 2016 – foundational GBM model
  + NoisET tool (Elhanati et al., 2021) for noise modeling in RepSeq data

1. **Reviewer’s Comments:**

* The model elegantly disentangles biological noise from genuine TCR dynamics
* Particularly powerful for understanding **age-related repertoire changes**
* Future extensions could include phenotype-resolved (e.g., naive/memory/CD4/CD8) modeling, or incorporation of single-cell data
* Could be paired with clustering tools to track **antigen-specific clone sets**
* Adding SHM modeling and BCR support would broaden utility

1. **Simulation Level / Output Granularity:**

* Clone-level and population-level dynamics
* Provides both summary parameters and posterior distributions per clone

1. **Visualization:**

* Scatter plots of clone frequencies across time
* Power-law distribution fits
* Turnover time vs. age plots
* Fold-change distributions
* Persistence probability curves

**1. Simulation as a Bridge Between Data and Mechanism in Autoimmunity**

This article exemplifies how **longitudinal TCR-seq data** can be paired with **stochastic modeling** to uncover mechanisms underlying immune homeostasis—even in the absence of explicit immune challenge. While the study focuses on healthy individuals, its approach is a **template for autoimmune modeling**, where tracking clone persistence or stochastic fluctuations can reveal flare predisposition.

**Relevance to Autoimmunity:**

* Autoimmune diseases are marked by long-term clonal dynamics and sporadic flares.
* The inferred **turnover parameters (τ and θ)** could help detect dysregulated turnover or persistence of autoreactive clones.
* The **noise-aware inference** helps distinguish true expansions from measurement error—critical in diagnosing autoimmune relapses.

**2. Benchmarking Crisis and the Need for Gold Standards**

This study sets a strong **benchmarking example** by:

* Using **synthetic data** to validate inference.
* Quantifying **sampling noise** using replicates.
* Using **NoisET** to model error and **compare naive vs full inference**.

**Implication for the Field:**

* The lack of standardized models for TCR dynamics in autoimmune conditions is a gap.
* The geometric Brownian motion framework inferred here can act as a **baseline null model** for autoimmunity vs healthy states.

**3. Simulation in the Era of Deep Learning and Multi-Omics**

While this article does **not apply deep learning**, its inferred parameters (τ, θ, α) could serve as **features for downstream ML models**.

**Suggestion for Future Work:**

* Combine clone dynamics with **single-cell transcriptomics** or **epitope-specific clustering** to build **multi-modal ML models**.
* Use the inferred clone-level dynamics to augment **digital twin simulations** in autoimmunity.

**4. Simulation for Personalized Immunology and Precision Medicine**

The study shows that **clone turnover slows with age** and that **clonal dynamics are frequency- and phenotype-dependent** (e.g., CD4 vs CD8).

**Toward Digital Twins:**

* Personalized models of τ and θ can **predict persistence or extinction of clones**, which is foundational to digital twin modeling of autoimmunity.
* Incorporating **patient-specific TCR-seq + age + phenotype + inferred parameters** could simulate individual immune trajectories.

**Clinical Translation:**

* Predicting the stability of pathogenic clones in autoimmune patients.
* Monitoring therapy response by modeling clone decline or resurgence.

**5. Critical Gaps and Future Directions: What the Field Needs Next**

**Identified Gaps:**

* No incorporation of **phenotype-resolved repertoire** (e.g., naive vs memory sorting).
* No direct modeling of **antigen specificity** or **SHM**.
* Assumes **clone dynamics are frequency-invariant**, but shows evidence against this.

**What the Field Needs:**

* Larger datasets with **sorted and barcoded single-cell multi-omics**.
* Integration with **simulation tools** (e.g., immuneSIM, AIRRSHIP) to validate dynamics.
* Application to **pathological repertoires** (autoimmune, cancer, infectious disease).

**Integration with "Individualized VDJ Recombination" Insight**

*Genome Res. 2021 Dec;31(12):2209–2224*

This article’s findings **complement** the Genome Res. study. While that study focused on **generation biases**, this one models **post-thymic clone dynamics**. Together, they illustrate:

* How **generation + dynamics = individual-specific repertoire evolution**
* The importance of **personalized modeling frameworks** in immune simulation
* The need for **benchmarks** that span both **generation (IGoR/OLGA)** and **dynamics (GBM models)**

**TULIP a Transformer based Unsupervised Language model for Interacting**

**Peptides and T-cell receptors that generalizes to unseen epitopes**

**Tool/Model Summary Template**

1. **Tool/Model Name:**  
   **TULIP** (Transformer-based Unsupervised Language Model for Interacting Peptides and T-cell Receptors)
2. **Reference(s):**  
   Barthelemy Meynard-Piganeau, Christoph Feinauer, Martin Weigt, Aleksandra M. Walczak, Thierry Mora (2023)  
   *bioRxiv* preprint. DOI: 10.1101/2023.07.19.549669
3. **Category:**  
   Deep Learning / Language Model; TCR–pMHC Binding Prediction; One-Class Classification; Unsupervised Learning
4. **Main Purpose/Function:**  
   Predicts TCR–epitope binding using an unsupervised transformer-based model trained solely on positive interaction data, avoiding biases from artificial negative data.
5. **Key Features:**
   * Transformer encoder-decoder architecture
   * Fully unsupervised learning (no need for negative examples)
   * Uses both complete and incomplete interaction data
   * Probabilistic output based on conditional likelihood
   * Generalizes to unseen epitopes
   * Repertoire mining capability
   * Resilient to negative sampling bias
6. **Input Requirements:**
   * Amino acid sequences of CDR3α, CDR3β, and epitopes
   * MHC class (as a categorical variable, not full sequence)
   * Can operate with missing chain(s) due to modular design
7. **Output:**
   * Conditional probabilities of binding (e.g., P(epitope|TCR, MHC))
   * Pointwise mutual information scores (as binding proxy)
   * Rank lists of TCRs likely to bind a given epitope
   * Spearman correlations with experimental EC50 values
8. **Validation/Evaluation:**
   * Compared to NetTCR-2.0, PanPep, ERGO2, DLpTCR
   * Evaluated on seen and **unseen epitopes**
   * ROC curves, AUC, and Spearman correlation with EC50
   * Robust to various negative sampling strategies
9. **Strengths:**
   * Avoids supervised learning biases
   * Can leverage incomplete data (single-chain, missing MHC, etc.)
   * Generalizes well to unseen epitopes
   * Outperforms or is competitive with state-of-the-art methods
   * Allows repertoire mining and mutation effect prediction
10. **Limitations:**

* Produces probabilistic scores, not binding constants (e.g., KD)
* Cannot assess risk of self-reactivity due to lack of data
* Requires careful interpretation of scores without hard threshold

1. **Applications in Autoimmunity:**

* Not yet directly applied to autoimmune datasets
* Highly suitable for autoimmunity-related studies (e.g., predicting cross-reactivity, neoantigen mimicry, repertoire skewing)

1. **Notable Citations/Use Cases:**

* Applied to predict TCRs binding p53 neoantigens (e.g., R175H in cancer)
* Used to assess mutation impact on epitope recognition (deep mutational scan)

1. **Reviewer’s Comments:**

* Innovative unsupervised approach avoids pitfalls of negative sampling
* Could benefit from integration with structural data for affinity estimation
* Promising tool for repertoire mining, especially in cancer and neoantigen settings
* Suggest including native self-epitope screening for autoimmunity assessment in future versions

1. **Simulation Level / Output Granularity:**

* Chain-level amino acid resolution (CDR3α, CDR3β, epitope)
* Probability distributions for binding sequences
* Position-level attention within sequences

1. **Visualization:**

* ROC and PPV curves (performance plots)
* AUC vs epitope distance (Levenshtein-based generalization plots)
* Rank vs repertoire size (repertoire mining results)
* Available via external plotting (BioRender, Python)

**1. Simulation as a Bridge Between Data and Mechanism in Autoimmunity**

TULIP is not just a prediction model—it's a **generative, unsupervised framework** that learns **TCR–epitope binding rules** directly from positive interaction data. This is especially important in autoimmunity where:

* Negative examples are rare or ambiguous (e.g., a TCR might bind self weakly but trigger pathology).
* Mechanisms are often unknown, yet **sequence-level signals** (e.g., CDR3 motifs, binding preferences) encode critical functional traits.

TULIP's **encoder–decoder structure** bridges observed data (TCR–epitope pairs) with latent binding mechanisms, creating a promising **simulation–mechanism interface**.

**Application Potential in Autoimmunity:**

* Can simulate TCRs likely to bind self-antigens
* Learn and model motif specificity for autoreactive clones
* Identify structural features that distinguish binding vs. tolerance

**2. Benchmarking Crisis and the Need for Gold Standards**

The article presents a **clear critique of current benchmarking practices**, especially in TCR binding prediction:

* **Bias from negative sample generation** in supervised models (e.g., random pairing)
* **Inflated performance metrics** due to test/train leakage or epitope imbalance
* Introduction of **rigorous testing protocols**, including:
  + UURA (Unseen Unconnected Random Association)
  + HRS (Healthy Repertoire Sampling)

TULIP, being **unsupervised**, is inherently robust to such sampling biases and offers a blueprint for **bias-resistant evaluation**.

**Call to Action:**

* The field urgently needs **standardized datasets**, **transparent splitting**, and **negative sampling protocols** to compare models meaningfully.

**3. Simulation in the Era of Deep Learning and Multi-Omics**

TULIP is at the forefront of integrating:

* **Transformer-based language modeling** (deep learning)
* **Multi-chain, multi-part inputs** (α-CDR3, β-CDR3, epitope, MHC)
* **Flexible input handling**, including incomplete data scenarios

Though not yet fused with transcriptomics or proteomics, the framework is **modular** enough to integrate omics features via embeddings.

**Future Direction:**

* Integrate expression (e.g., TCR clone expansion), methylation, or single-cell RNA-seq data to train **multi-modal immune representations**.
* Use generated repertoires in **virtual cohort simulations** or **neoepitope vaccine design**.

**4. Simulation for Personalized Immunology and Precision Medicine**

TULIP demonstrates strong **generalization to unseen epitopes**, critical for:

* **Patient-specific neoantigen response prediction**
* **Cross-reactivity detection**
* **Monitoring therapy response or autoantigen reactivity**

It also supports **repertoire mining**: finding known binders in a background of healthy TCRs, a step toward **immune digital twins**.

**Example in the Paper:**

* **p53 neoantigen (HMTEVVRHC)**: TULIP successfully recovers known TCRs even when trained without them.

**Digital Twin Perspective:**

* Use TULIP to simulate “what-if” scenarios (e.g., mutation-induced binding gain/loss).
* Predict autoimmune flares or immune escape in silico using longitudinal patient data.

**5. Critical Gaps and Future Directions: What the Field Needs Next**

**Identified Limitations in the Paper:**

* TULIP models *binding probability*, not binding affinity constants (e.g., no EC50 calibration).
* Lack of structural data—only sequence-level modeling.
* Still limited epitope diversity in training (real-world challenge).
* Cannot predict off-target binding (e.g., cross-reactivity with self-peptides) without large-scale validated non-binders.

**Recommendations for the Field:**

* Combine with **AlphaFold/structural modeling** for 3D-aware prediction.
* Develop **simulation pipelines** that include repertoire generation (IGoR/OLGA), dynamics (GBM), and binding (TULIP-like models).
* Integrate with **multi-omics** to build fully context-aware predictors.

**Use with Genomic Data (from VDJ recombination studies like Genome Res. 2021):**

* Account for **individual VDJ biases** (genetic and epigenetic) in training sets
* Generate **individualized TCR distributions**, boosting simulation realism and personalizability

**Summary**

TULIP sets a new standard for **unsupervised, bias-resistant TCR–epitope interaction modeling**. Its flexible architecture, strong generalization, and applicability to repertoire mining make it a foundational building block for:

* **Mechanistic insights** in autoimmune T cell recognition
* **Benchmarking and fair model evaluation**
* **Deep learning–based simulation and data augmentation**
* **Patient-specific immune modeling and digital twin applications**

**TAPIR: a T-cell receptor language model for predicting rare and novel targets**

**1. Tool/Model Name:**

**TAPIR** (T-cell receptor and Peptide Interaction Recognizer)

**2. Reference(s):**

Fast E, Dhar M, Chen B. *TAPIR: a T-cell receptor language model for predicting rare and novel targets.* bioRxiv. 2023.  
https://doi.org/10.1101/2023.09.12.557285

**3. Category:**

* Deep Learning
* TCR-pMHC Interaction Prediction
* TCR Repertoire Modeling
* In Silico TCR Design

**4. Main Purpose/Function:**

TAPIR predicts whether a given T-cell receptor (TCR) will interact with a peptide-MHC (pMHC) target. It is particularly designed to generalize to **novel and rare targets** not seen during training, addressing a major challenge in TCR-based immunotherapy discovery.

**5. Key Features:**

* Deep CNN-based two-tower architecture
* Handles **paired and unpaired** TCR chains
* **Target sequence-agnostic** prediction (i.e., can predict even when target is unknown)
* Supports **missing inputs** (e.g., masked V/J/CDR3/MHC)
* Compatible with **TCRα, TCRβ, or both**
* Accepts **MHC alleles** and peptide sequences as input
* Capable of **in-silico TCR design** when extended with a generative module
* Open-access server for academic use: https://vcreate.io/tapir

**6. Input Requirements:**

* TCR sequences (α/β chains, CDR3, V/J genes)
* Optional: peptide sequences and/or MHC allele
* Masking tokens ('X') allowed for missing elements
* Works with both public datasets (VDJdb) and proprietary functional datasets

**7. Output:**

* Predicted **interaction score** (probability of TCR–target binding)
* Optionally: AI-designed TCRs (with generative extension)
* AUROC for evaluation tasks
* Visualizations (in paper) for performance comparison and amino acid importance

**8. Validation/Evaluation:**

* Benchmarked against **DeepTCR, TCRAI, NetTCR**
* Tested on **held-out known targets** and **zero-shot novel targets**
* **Functional validation** against real-world cancer antigens (e.g., PIK3CA H1047L)
* Designed TCRs were **experimentally validated** via NFAT and CD69 assays

**9. Strengths:**

* State-of-the-art **zero-shot** prediction capability
* **Flexible input format** (handles missing data)
* Demonstrated **functional discovery** of novel anti-cancer TCRs
* Capable of **designing** TCRs in silico with validation
* Publicly accessible web server and datasets

**10. Limitations**

* Performance can still vary across novel targets due to limited training data
* Not yet a full replacement for wet lab screening; best used as **complementary**
* Requires more **diverse functional training data** for further improvement

**11. Applications in Autoimmunity:**

* Not directly tested in autoimmune contexts
* However, capable of modeling TCR-target interactions and could be extended to self-antigens (e.g., proinsulin test cases mentioned)
* Suitable for future autoimmune disease studies due to generalizable design

**12. Notable Citations/Use Cases:**

* Used to discover and validate a **novel TCR against PIK3CA H1047L**, a common cancer neoantigen
* Designed **6 new TCRs** against the influenza peptide GILGFVFTL; 3 bound target, 2 showed activation

**13. Reviewer’s Comments:**

* TAPIR is among the **most generalizable TCR predictors** to date
* The **combination of CNN architecture + data augmentation** is highly innovative
* Integration with structural modeling (e.g., AlphaFold2) in validation strengthens confidence
* Adding support for autoimmune-specific datasets and structural constraints could enhance utility

**14. Simulation Level / Output Granularity:**

* Granularity: **Single TCR–target pair-level interaction**
* Score ranges from 0 to 1, suitable for ranking and filtering candidates

**15. Visualization:**

* AUC and AUROC plots for performance comparisons
* Violin plots for binding scores
* Tetramer binding and activation assays (NFAT, CD69)
* Structural visualizations using **AlphaFold2 + PyMOL** to map amino acid importance in binding

**1. Simulation as a Bridge Between Data and Mechanism in Autoimmunity**

**TAPIR** acts as a bridge between **high-throughput AIRR-seq data** and **functional TCR-antigen mechanisms** by learning generalizable sequence–function relationships.

* It uses both **paired and unpaired TCR data**, integrates **MHC and peptide inputs**, and handles incomplete/masked information.
* Demonstrates that deep learning models can **predict activation**, not just binding, enabling a more **mechanistic interpretation** of immune recognition.

**Autoimmunity Implication**:  
This opens the door to simulating and interpreting TCR cross-reactivity with **self-antigens**, a core issue in autoimmune disease. TAPIR could simulate whether disease-associated TCRs are likely to activate against mutated or mimicry peptides.

**2. Benchmarking Crisis and the Need for Gold Standards**

The authors recognize and **explicitly address the benchmarking issue**:

* Compare against prior models (NetTCR, DeepTCR, TCRAI) using standardized AUC metrics.
* Evaluate performance on **42 novel, unseen targets** (zero-shot prediction), rare in current literature.
* Introduce **functional activation benchmarks**, not just binding.

**Call to the Community**:  
This work emphasizes the **urgent need for functionally annotated TCR datasets**. It sets a new precedent for benchmarking not only accuracy but **clinical relevance (e.g., NFAT/CD69 expression)**.

**3. Simulation in the Era of Deep Learning and Multi-Omics**

TAPIR is a **deep learning-first** model:

* Built on **CNN-based dual encoders** (one for TCR, one for target/MHC).
* Trained with **extensive data augmentation** (masking V/J, missing chains, target noise).
* Capable of generalizing to **targets never seen in training**.

**Integration Potential**:

* TAPIR already works with **multi-format inputs** (e.g., just β chain + MHC, or α/β + full target).
* Could easily be expanded to incorporate **transcriptomic phenotypes** (e.g., exhaustion markers) or **antigen presentation models** (NetMHCpan).

**4. Simulation for Personalized Immunology and Precision Medicine**

TAPIR has direct implications for **digital twin construction** and **precision TCR therapy**:

* Used to identify and validate a **novel anti-PIK3CA TCR**, with **greater specificity and potency** than the previously known therapeutic clone (C-66).
* Extended with a **generative model** that **designs new TCRs**, validated in both **tetramer binding** and **activation assays** (NFAT, CD69).

**Precision Medicine Vision**:

* Can score a patient's entire TCR repertoire against a panel of autoantigens, predicting risk or **flare reactivation**.
* Forms the backbone for **in silico clinical trial simulations**, similar to digital twins in oncology.

**5. Critical Gaps and Future Directions: What the Field Needs Next**

**Unmet Needs Highlighted by TAPIR**:

* **Functional training data** remains limited and proprietary.
* **Structural integration (e.g., AlphaFold-based co-crystal modeling)** is missing, though feasible.
* TAPIR predicts **interactivity score**, not true binding affinity (e.g., KD).

**Future Directions**:

* **Public, standardized datasets** with activation readouts.
* Fusion with **3D structural predictors** for better epitope modeling.
* Cross-modal learning: integrate **presentation (MHC), expression, mutation load**.
* Simulation of **individualized repertoires** using generation models (e.g., IGoR + TAPIR).

**Integration with "Individualized VDJ Recombination" (Genome Res. 2021)**

**Reference**: Genome Research, 2021; 31(12):2209–2224. DOI: 10.1101/gr.275373.121

TAPIR, though trained on pooled TCR datasets, can be extended to **individual-specific prediction** by:

* Integrating **personalized generation models** (from OLGA or IGoR) to simulate a patient’s potential TCR space.
* Scoring these against **autoantigens or neoantigens** to predict risk.

This would form the foundation of **fully individualized immuno-digital twins**, where **generation + prediction + dynamics = personal immunological map**.

**AIRRSHIP: simulating human B cell receptor repertoire sequences**

**Tool/Model Summary Template – AIRRSHIP**

1. **Tool/Model Name:**  
   **AIRRSHIP** (Adaptive Immune Receptor Repertoire Simulation of Human Immunoglobulin Production)
2. **Reference(s):**  
   Sutherland C, Cowan GJM.  
   *AIRRSHIP: simulating human B cell receptor repertoire sequences*.  
   *Bioinformatics*, 2023; 39(6):btad365.  
   [DOI: 10.1093/bioinformatics/btad365](https://doi.org/10.1093/bioinformatics/btad365)  
   GitHub: <https://github.com/Cowanlab/airrship>  
   Docs: <https://airrship.readthedocs.io>
3. **Category:**
   * Full Repertoire Simulator
   * SHM Model
   * Benchmarking Tool
4. **Main Purpose/Function:**  
   To generate **synthetic B cell receptor (BCR) sequences** that accurately replicate real human immunoglobulin repertoire features.  
   AIRRSHIP is designed to provide **ground-truth datasets** for benchmarking annotation, alignment, and immune repertoire analysis tools.
5. **Key Features:**
   * Implemented in **Python**, no extra dependencies
   * Generates **realistic full-length BCR sequences**
   * Simulates:
     + **VDJ recombination** with trimming and insertions
     + **Somatic hypermutation (SHM)** at sequence and position level
   * Uses **Markov model** for realistic nucleotide insertions
   * Allows user control over: VDJ usage, indel distributions, SHM rates
   * Outputs **AIRR-compliant** TSV + FASTA files
   * Fast generation (e.g., 10K sequences in <70 sec)
   * Supports benchmarking of IgBLAST, MiXCR, IMGT/HighV-QUEST, etc.
6. **Input Requirements:**
   * No external input needed for default use
   * Optional: user-provided reference gene usage, trimming, or SHM models
   * Parameters for simulation settings (gene usage, mutation rates, etc.)
7. **Output:**
   * **FASTA** file of synthetic BCR sequences
   * **TSV** annotation file (follows AIRR rearrangement schema)
   * Metadata includes: VDJ gene assignments, mutation sites, trimming lengths, junction insertions, SHM statistics
8. **Validation/Evaluation:**
   * Compared to real BCR datasets using **sumrep** (for CDR3 length, VDJ usage, junction insertions, SHM pattern)
   * Correlation of SHM patterns with real data: **r = 0.97**
   * Outperforms or matches **immuneSIM**, **IMPlAntS**, and **partis** in realism
   * Accurate amino acid usage, GRAVY score, aliphatic index, Atchley factor scores
   * More realistic junction insertions than other tools
9. **Strengths:**
   * High **realism and speed**
   * Minimal setup: runs out-of-the-box
   * **Highly customizable** for benchmarking scenarios
   * AIRR-compliant and tool-compatible output
   * Uses broad and deep reference datasets (>11 million sequences)
   * Markov insertion model closely replicates real junctional diversity
10. **Limitations:**

* Currently limited to **human heavy chain BCRs**
* No TCR or paired chain simulation
* No built-in GUI
* Not a sequence annotation or analysis tool

1. **Applications in Autoimmunity:**

* Not directly applied to autoimmune disease datasets
* However, **ideal for simulating BCRs** with **disease-related SHM/motif patterns**
* Enables controlled experiments to test annotation performance in **autoimmune-like contexts**

1. **Notable Citations/Use Cases:**

* This is a **recent tool**; no independent citations yet
* Compared against: **immuneSIM**, **IMPlAntS**, **partis**
* Validated against IgD, IgM, IgG repertoires and mutation models

1. **Reviewer’s Comments:**  
   AIRRSHIP fills a critical gap in simulating high-quality, realistic human BCR repertoires for benchmarking. Its balance of realism, flexibility, and speed make it highly useful. Extending to TCRs or light chains, and allowing SHM-motif implantation, would enhance its applicability.
2. **Simulation Level / Output Granularity:**

* **Sequence-level** simulation with full VDJ recombination, SHM, and productivity annotations
* Granular tracking of mutation and recombination events
* AIRR-formatted metadata

1. **Visualization:**

* Not built-in
* Compatible with R (e.g., **sumrep**) or Python for:
  + VDJ usage plots
  + SHM heatmaps
  + Insertion length distributions
  + Sequence composition (GRAVY, aliphatic index)

**1. Simulation as a Bridge Between Data and Mechanism in Autoimmunity**

**AIRRSHIP** is purpose-built to create high-fidelity synthetic BCR repertoires that closely mirror real human immunoglobulin sequences. By reproducing detailed mechanisms such as:

* **VDJ recombination**
* **Junctional trimming**
* **Position-specific nucleotide insertion (via Markov chains)**
* **Context-dependent SHM**

…it connects **mechanistic immunobiology** with **data-driven analysis**.

**Autoimmunity Connection:**

* Simulated repertoires can mimic autoantibody diversity and mutation patterns.
* By adjusting SHM intensity or junctional variability, AIRRSHIP can simulate **pathological B cell populations** found in autoimmune diseases (e.g., SLE, RA).

**2. Benchmarking Crisis and the Need for Gold Standards**

The authors **explicitly address the benchmarking gap**:

* Current tools (IgBLAST, MiXCR, IMGT) often **disagree on basic annotations** (e.g., VD insertion length).
* AIRRSHIP includes a **TSV file with full ground truth** for every simulated sequence, enabling **direct, quantitative benchmarking**.

**Takeaway**: AIRRSHIP is more than a simulator—it’s a **gold standard dataset generator**. The paper even provides a **use case comparing tool errors** across different trimming conditions.

**Roadmap Suggestion**:

* Community-wide use of AIRRSHIP to validate repertoire analysis tools.
* Establish benchmarking challenges with defined “truth sets”.

**3. Simulation in the Era of Deep Learning and Multi-Omics**

While AIRRSHIP is not itself a deep learning tool, it **enables DL**:

* By generating **large-scale, realistic repertoires** (~1M sequences in <2h), it provides training data for machine learning and deep learning models.
* Outputs are **AIRR-compliant**, allowing easy integration with downstream omics pipelines.

**Use Case**:

* Train autoencoder-based models to detect **mutation hotspots or clonal convergence** in autoimmune BCR repertoires using AIRRSHIP data.

**Future Integration Needed**:

* Include transcriptomic features (e.g., isotype, activation state).
* Pair with scRNA-seq simulations like **Echidna** for joint receptor-transcriptome modeling.

**4. Simulation for Personalized Immunology and Precision Medicine**

AIRRSHIP allows user-defined control over:

* **VDJ usage**
* **Insertion/deletion distributions**
* **SHM frequency and motif-specific mutation rates**

This makes it ideal for **personalized immunology**, where we simulate repertoires for:

* **Autoimmune-prone individuals**
* **Monogenic B cell deficiencies**
* **Response to B-cell-targeting therapies**

**Toward Digital Twins:**

* AIRRSHIP can simulate a virtual patient’s naïve or memory BCR repertoire under various parameters (e.g., different V gene usage biases).
* When combined with models of clonal selection and antigen affinity, this forms a scaffold for **immune digital twin construction**.

**5. Critical Gaps and Future Directions: What the Field Needs Next**

**Gaps:**

* AIRRSHIP currently supports **only BCR heavy chains** and **human sequences**.
* No modeling of **antigen-driven selection** or **structural affinity**.
* No direct support for **paired-chain simulation** or **longitudinal repertoire dynamics**.

**Call to Action**:

* Extend AIRRSHIP to:
  + Paired heavy/light chain generation
  + TCRαβ support
  + Modeling clonal expansion, deletion, and antigen affinity
* Merge with frameworks like **TAPIR** (for TCR affinity prediction) or **immuneSIM** (for murine models and class-switched data)

**Tying to VDJ Individualization (Genome Res. 2021)**:  
AIRRSHIP draws from over **11 million BCR sequences from 380 individuals**, making it already **aware of individual VDJ recombination variation**—a rare strength.

Future iterations could:

* Accept **personalized recombination profiles** inferred from real data
* Simulate immune landscapes for specific genotypes or clinical contexts

**High-throughput immune repertoire analysis with IGoR**

**Tool/Model Summary Template – IGoR**

1. **Tool/Model Name:**  
   **IGoR** (Inference and Generation of Repertoires)
2. **Reference(s):**  
   Marcou Q, Mora T, Walczak AM.  
   *High-throughput immune repertoire analysis with IGoR*.  
   Nature Communications, 2018; 9:561.  
   https://doi.org/10.1038/s41467-018-02832-w
3. **Category:**
   * V(D)J Recombination Inference
   * SHM (Somatic Hypermutation) Modeling
   * Sequence Generator
   * Probabilistic Annotation Tool
4. **Main Purpose/Function:**  
   To infer the statistical rules of immune receptor generation (TCR/BCR) through probabilistic modeling of V(D)J recombination and somatic hypermutation, and to generate synthetic sequences with realistic properties for diagnostic and research purposes.
5. **Key Features:**
   * Probabilistic scenario assignment for recombination
   * Context-dependent SHM modeling using PWMs
   * Modular and flexible (customizable recombination graphs)
   * Sequence generation mode with learned statistics
   * Expectation–Maximization algorithm for parameter inference
   * Bayesian network structure for recombination event dependencies
   * Hypermutation co-localization detection
   * Handles both TCR and BCR, and DNA or RNA input
6. **Input Requirements:**
   * Pre-processed unique sequences (from cDNA or gDNA)
   * Germline gene segment databases (e.g., IMGT)
   * Optional: read quality control and sequence filtering beforehand
7. **Output:**
   * Ranked recombination scenarios per sequence with likelihoods
   * Generation probabilities
   * SHM motif PWMs
   * Synthetic repertoires
   * Statistical summaries of recombination and mutation patterns
8. **Validation/Evaluation:**
   * Validated on synthetic data with known statistics
   * Compared to MiXCR and Partis: IGoR had ~2× better accuracy in identifying full recombination scenarios
   * Robust across different sequencing platforms and individual variation
   * High correlation (r = 0.99) between inferred and true distributions
9. **Strengths:**
   * Accurately captures V(D)J recombination diversity
   * Handles degenerate recombination paths
   * Context-aware SHM inference
   * Highly interpretable and reproducible
   * Applicable across species with appropriate germline DB
   * Suitable for both analysis and simulation
10. **Limitations:**

* No explicit GUI (command-line only)
* Limited performance on sequences with very high SHM unless models are trained on naive data
* Computationally intensive for large datasets due to full scenario enumeration
* Requires good germline gene annotations
* Indels not explicitly modeled (although partially handled)

1. **Applications in Autoimmunity:**

* Not directly evaluated in autoimmune disease studies
* Suitable for identifying convergent recombination and public clonotypes, which are often relevant in autoimmune settings
* Can distinguish between generation and selection processes, aiding in autoimmune sequence interpretation

1. **Notable Citations/Use Cases:**

* Elhanati et al., 2016 – repgenHMM development comparison
* Pogorelyy et al., 2017 – convergent recombination in TCR repertoires
* Widely used for baseline generation probability calculations

1. **Reviewer’s Comments:**  
   IGoR represents a major step forward in probabilistic annotation and generation of immune receptor repertoires. Its strength lies in robust modeling and flexibility. Future development could benefit from GUI integration, indel modeling, and extension to paired chain data. Particularly valuable in immunodiagnostics, vaccine design, and repertoire selection analysis.
2. **Simulation Level / Output Granularity:**

* Clone-level sequences with detailed scenario probability
* Sequence-level generation probability and SHM annotation
* PWM-based mutation motif profiles for V, D, and J

1. **Visualization:**

* Not built-in; however, IGoR outputs can be visualized with external tools (e.g., R, Python)
* Common visualizations: mutation heatmaps, insertion/deletion distributions, motif logos, co-localization plots

**1. Simulation as a Bridge Between Data and Mechanism in Autoimmunity**

**IGoR** (Inference and Generation of Repertoires) builds a robust computational bridge between **high-throughput AIRR-seq data** and **mechanistic understanding** of receptor generation by:

* Modeling **probabilistic recombination scenarios** (V(D)J, insertions, deletions).
* Simulating **context-dependent SHM** in B cells.
* Providing full distributions over **generation probabilities**, essential for interpreting rare sequences such as autoreactive clones.

**Autoimmunity Relevance**:

* Helps dissect whether a disease-associated receptor is **public (convergently generated)** or **functionally selected**.
* Informs **baseline generation probabilities**, crucial for distinguishing autoreactive clones from statistical noise.

**2. Benchmarking Crisis and the Need for Gold Standards**

IGoR directly tackles the **benchmarking problem**:

* Provides a **ground-truth generator** to create synthetic repertoires with known recombination parameters.
* Evaluates performance using **KL-divergence** and **scenario ranking** rather than only best-match assignment.
* Shows that **most tools misassign recombination scenarios**, while IGoR captures degeneracy explicitly.

**Proposed Benchmarking Roadmap**:

* Use IGoR-generated data to validate annotation tools (e.g., MiXCR, Partis).
* Promote **probabilistic benchmarking** over deterministic assignment.
* Develop shared “truth-set” repertoires for evaluating repertoire analysis tools.

**3. Simulation in the Era of Deep Learning and Multi-Omics**

Though IGoR is not a deep learning model, it **enables deep learning** by:

* Generating **millions of realistic receptor sequences** for training.
* Offering **feature-rich probabilistic annotations** (e.g., SHM scores, generation probabilities).

**Integration Potential**:

* Pair with **DL models** for repertoire classification (e.g., antigen prediction).
* Simulate repertoires with **controlled biases or mutations**, aiding DL interpretability.
* Extend to **multi-omics integration**, e.g., using transcriptome-guided mutation rates.

**4. Simulation for Personalized Immunology and Precision Medicine**

IGoR supports **individual-specific modeling** by:

* Allowing inference of **personal recombination statistics** from naive, out-of-frame sequences.
* Highlighting that **insertion and deletion patterns are universal**, but **gene usage varies** by individual.

**Toward Digital Twins**:

* IGoR can simulate the **likely sequence space** of a specific patient’s immune system.
* Incorporating inferred SHM models allows simulation of **affinity-matured repertoires**, including potentially autoreactive BCRs.
* Provides **generation probability priors** for distinguishing disease-specific clones from expected background.

**5. Critical Gaps and Future Directions: What the Field Needs Next**

**Limitations Identified**:

* Lacks structural modeling (e.g., no peptide binding/affinity prediction).
* Hypermutation modeling doesn’t include **insertions/deletions** or **chromatin effects**.
* Simulates only **germline–derived mutations**; doesn’t cover clonal expansion or dynamics.

**Call to Action**:

* Expand to **long-read data**, paired chains, and non-human repertoires.
* Integrate with structural/functional predictors (e.g., TAPIR, TCRAI).
* Combine with models of **selection, expression, and evolution** for full immune system simulation.

**Connection to VDJ Individualization (Genome Res. 2021)**

*Genome Res. 2021;31(12):2209–2224. DOI: 10.1101/gr.275373.121*

The Genome Research paper shows that **VDJ recombination is individual-specific**, modulated by genetics and non-genetic factors.

**IGoR Implications**:

* Already accounts for this variability: allows training on **individual repertoires**.
* Highlights gene usage differences and universal insertion/deletion statistics.
* Supports future simulation of **patient-specific immune landscapes**.

**immuneSIM: tunable multi-feature simulation of B- and T-cell receptor repertoires for immunoinformatics benchmarking**

**Tool/Model Summary Template – immuneSIM**

1. **Tool/Model Name:**  
   **immuneSIM**
2. **Reference(s):**  
   Weber CR, Akbar R, Yermanos A, Pavlović M, Snapkov I, Sandve GK, Reddy ST, Greiff V.  
   *immuneSIM: tunable multi-feature simulation of B- and T-cell receptor repertoires for immunoinformatics benchmarking*.  
   *Bioinformatics*, 2020, 36(11):3594–3596.  
   DOI: 10.1093/bioinformatics/btaa158  
   GitHub: <https://github.com/GreiffLab/immuneSIM>  
   Docs: https://immuneSIM.readthedocs.io
3. **Category:**
   * Full Repertoire Simulator
   * SHM Modeling
   * Benchmarking Tool
   * Sequence Simulation Tool
4. **Main Purpose/Function:**  
   To generate fully annotated synthetic immune repertoires (TCR/BCR, human/mouse, single/paired chains) that replicate **native-like or aberrant immune repertoires** with tunable immunological features for the purpose of **benchmarking immunoinformatics tools**.
5. **Key Features:**
   * R package; open-source
   * **Tunable** simulation of V(D)J recombination and SHM
   * Supports both **BCR and TCR**, single and paired chain
   * Generates **native-like** and **aberrant** repertoires
   * **Traceable annotations** for all simulation events
   * Motif implantation, codon replacement, and sequence graph modification
   * Parallelization support for large-scale simulations
6. **Input Requirements:**
   * User-defined simulation parameters (e.g., V(D)J gene usage, insertion/deletion rates, SHM settings, motif implantation)
   * Optional reference datasets (for native-like repertoires)
7. **Output:**
   * Annotated simulated receptor sequences (full-length)
   * Event log for each simulated sequence (gene usage, insertions, deletions, SHM, etc.)
   * Sequence-level metadata (species, chain type, productivity, etc.)
   * Compatible with downstream analysis tools
8. **Validation/Evaluation:**
   * Simulated repertoires compared with experimental datasets (CDR3 length, VDJ usage, amino acid distributions, gapped k-mers)
   * 99% of sequences called **productive and in-frame** by IMGT/HighV-Quest
   * High overlap in annotations (V: >97%, J: >97%, D: ~60%)
   * Spearman correlation of >0.985 for gene usage; high similarity for motifs and abundance
9. **Strengths:**
   * Highly **flexible and configurable** simulator
   * Excellent validation against real data
   * Enables **benchmarking across multiple dimensions** (gene usage, diversity, SHM, motifs, etc.)
   * Provides **ground truth** for algorithm development
   * Supports **motif implantation** for ML tool testing
10. **Limitations:**

* Requires knowledge of R and scripting
* No GUI interface
* Limited to simulation (not for inference or analysis of real repertoires)
* D gene annotation match (~60%) lower than V/J (due to biological variability)

1. **Applications in Autoimmunity:**

* Not directly applied to autoimmune disease, but highly suitable
* Can simulate **public vs. private clones**, implanted **disease-related motifs**, and repertoire perturbations
* Useful for **testing diagnostic algorithms** and ML methods in autoimmune settings

1. **Notable Citations/Use Cases:**

* Greiff et al., 2017 – Learning repertoire features
* Emerson et al., 2017 – Machine learning classifier testing
* Glanville et al., 2017 – Motif-specific simulation for clustering studies

1. **Reviewer’s Comments:**  
   immuneSIM is an essential benchmarking tool for the immunoinformatics community. Its tunable design and validation against real data make it ideal for testing and developing machine learning methods, sequence annotation pipelines, and repertoire comparison frameworks. Future directions could include GUI integration and support for paired-chain selection models.
2. **Simulation Level / Output Granularity:**

* Full-length **V(D)J receptor sequences** (nucleotide and amino acid)
* Chain-level metadata and clone abundances
* Granular event-level tracking for every sequence (gene usage, SHM, etc.)

1. **Visualization:**

* Not built-in, but outputs compatible with standard R/Python tools
* Suggested visualizations:
  + CDR3 length distribution
  + V/D/J usage barplots
  + Heatmaps of amino acid frequency
  + k-mer co-occurrence graphs
  + Similarity network diagrams (after repertoire simulation)

**1. Simulation as a Bridge Between Data and Mechanism in Autoimmunity**

**immuneSIM** connects mechanistic processes of adaptive immunity (e.g., VDJ recombination, SHM) with synthetic AIRR-seq data. It enables the simulation of:

* **Native-like repertoires** that replicate experimental datasets.
* **Aberrant repertoires** representing autoimmune states or disease-specific perturbations.

**Autoimmunity Application**:

* Can mimic **clonal expansion**, **hypermutation profiles**, or **sequence motif patterns** observed in autoreactive B or T cells.
* Simulated datasets allow probing how **immune dysregulation** translates into repertoire-level signatures.
* Useful for evaluating how well ML tools can detect **disease-associated motifs** or **public/private clone structures**.

**2. Benchmarking Crisis and the Need for Gold Standards**

The paper directly tackles the **lack of standardized benchmarking** in immunoinformatics:

* Provides **fully annotated ground truth** for each sequence (gene usage, SHM, insertions, deletions).
* Shows >97% agreement with IMGT for V/J gene calls and ~60% for D genes—demonstrating its realism for testing annotation tools.

**Significance**:

* immuneSIM acts as a **universal benchmarking dataset generator**.
* It allows stress-testing of ML and alignment tools across a range of conditions, including "diseased" or noisy repertoires.

**3. Simulation in the Era of Deep Learning and Multi-Omics**

While immuneSIM is an **R-based rule-driven simulator**, it:

* Provides **millions of realistic, annotated sequences** for deep learning model training.
* Enables **motif implantation** for interpretability testing in neural networks (e.g., does the model learn what you implanted?).

**Integration Potential**:

* Combine with **multi-omics simulators** (e.g., Echidna) for transcriptome + receptor data.
* Use in pre-training models like **TAPE**, **ESM**, or **TULIP** using immuneSIM-generated repertoires with known motifs or clone frequencies.

**4. Simulation for Personalized Immunology and Precision Medicine**

immuneSIM enables **personalization** by allowing:

* User-defined **germline gene usage**, **mutation rates**, and **motif frequencies**.
* Simulation of **paired chains** (e.g., TCRα–β or IgH–L), relevant for **patient-specific immune profiles**.

**Toward Digital Twins**:

* Simulate repertoires from individuals with known germline polymorphisms or clinical histories.
* Implant disease-specific motifs or simulate autoimmune skewing of clonal abundance.
* Foundation for **personalized vaccine response prediction** or **autoimmune flare simulation**.

**5. Critical Gaps and Future Directions: What the Field Needs Next**

**Current Gaps**:

* No modeling of **antigen specificity** or **binding affinity**.
* No explicit simulation of **longitudinal immune dynamics** or **tissue-specific repertoires**.
* No integration with **structural models** (e.g., antibody–antigen docking).

**Call to Action**:

* Combine with tools like **OLGA/IGoR** (generation probabilities), **TAPIR/TULIP** (binding prediction), or **AIRRSHIP** (SHM realism).
* Extend to simulate repertoires based on **individual-specific VDJ recombination rules** (as shown in Genome Res. 2021).

**Tie-In with Individualized VDJ Recombination (Genome Res. 2021)**

“VDJ recombination rules are genetically and environmentally modulated.”

**Relevance**:

* immuneSIM currently uses **experimental distributions**, but can incorporate **individual-specific gene usage and SHM biases**.
* This is key for modeling **immune variability in monozygotic twins**, or in autoimmune-prone individuals.

**Future Work**:

* Integrate **individual genotypes** or **VDJ usage from real patients** to create personalized immune simulations.
* This could feed into **digital twin pipelines** for in silico immunopathology modeling.

**Modeling and predicting the overlap of B- and T-cell receptor repertoires in healthy and SARS-CoV-2 infected individuals**

**1. Tool/Model Name:**

**Pgen + SONIA / soNNia** (Statistical Repertoire Sharing Framework)

**2. Reference(s):**

Ruiz Ortega M, Spisak N, Mora T, Walczak AM (2023).  
*Modeling and predicting the overlap of B- and T-cell receptor repertoires in healthy and SARS-CoV-2 infected individuals*. PLoS Genet 19(2): e1010652.  
https://doi.org/10.1371/journal.pgen.1010652

**3. Category:**

* **Statistical Modeling**
* **Repertoire Sharing Analysis**
* **Selection Modeling**
* **Diagnostics Tool**

**4. Main Purpose/Function:**

* To model, quantify, and predict the overlap of immune receptor repertoires (BCRs and TCRs) across individuals.
* To identify public clonotypes and assess the influence of convergent recombination and selection.
* To detect condition-associated (e.g., COVID-19) clonotypes using deviations from statistical expectation.

**5. Key Features:**

* Probabilistic modeling of V(D)J recombination using **IGoR**.
* Selection modeling using **SONIA** (linear) and **soNNia** (neural network).
* Incorporates convergent recombination and convergent selection.
* Can distinguish public clonotypes arising from chance vs. antigen-driven selection.
* Diagnostic capability using likelihood and logistic regression classifiers.
* Supports both BCR and TCR repertoires.

**6. Input Requirements:**

* High-throughput sequencing data of immune repertoires (BCR or TCR).
* Annotation of V, D, J gene usage and CDR3 region.
* Productive and non-productive sequence separation.
* Optional metadata: age, infection status, HLA type (not required but relevant for future improvements).

**7. Output:**

* Predicted sharing spectrum (distribution of public vs. private clonotypes).
* Identification of significantly overshared (disease-associated) clonotypes.
* Diagnostic scores (likelihood ratios or logistic regression probabilities).
* Gene usage analysis and network clustering of clonotypes.

**8. Validation/Evaluation:**

* Validated on repertoires from:
  + 10 healthy donors (BCRs, IgM and IgG).
  + 666 healthy donors (TCRβ).
  + 44 SARS-CoV-2+ BCR repertoires.
  + 1414 SARS-CoV-2+ TCRβ repertoires.
* Performance evaluated against known SARS-CoV-2 antibodies (BCRs) and MIRA-assayed TCRs.
* Comparison with immuneSIM and vampire tools; showed higher accuracy and lower runtime.

**9. Strengths:**

* Accurate prediction of receptor sharing using a well-grounded statistical model.
* Captures both generation and selection processes.
* Identifies public disease-associated clonotypes without requiring epitope knowledge.
* Applicable to both naive and memory repertoires.
* Can inform diagnostics and vaccine design.

**10. Limitations:**

* Requires large, high-quality repertoires for robust modeling.
* For TCRs, HLA-dependent selection is not explicitly modeled.
* Model underestimates sharing in highly antigen-driven repertoires without manual correction (via factor *q*).
* Lacks built-in visualization interface.

**11. Applications in Autoimmunity:**

* Not directly applied to autoimmunity in this study.
* Framework is suitable for identifying condition-associated public clonotypes in autoimmune diseases.
* Has previously been discussed in the context of ankylosing spondylitis (cited in discussion).

**12. Notable Citations/Use Cases:**

* [16] Emerson RO et al. 2017, Nat Genet – CMV status and TCR sharing.
* [70] Snyder TM et al. 2020 – MIRA assay of SARS-CoV-2-specific TCRs.

**13. Reviewer’s Comments:**

* Excellent integration of generative and selection models.
* The use of *q* as a convergent selection factor is elegant but could be refined to include personalized variables like HLA type.
* Future extensions may consider integrating temporal/longitudinal data and paired chains (α/β or heavy/light).

**14. Simulation Level / Output Granularity:**

* Amino acid sequence level (CDR3).
* Clone-level resolution (clonotypes).
* Sharing statistics per individual and cohort.

**15. Visualization:**

* Sharing number distributions (histograms).
* Pmodel vs. Pdata scatter plots.
* V and J gene usage bar plots.
* Network graphs (Levenshtein-based clustering).
* Sequence logos for motifs within clusters.
* ROC curves for diagnostic performance.

**1. Simulation as a Bridge Between Data and Mechanism in Autoimmunity**

This article showcases how **statistical models of immune receptor generation and selection** bridge mechanistic understanding (e.g., VDJ recombination, selection pressures) and real-world immune repertoire data (e.g., AIRR-seq, COVID-19 repertoires).

* **Tools used**:
  + IGoR for modeling V(D)J recombination (Pgen)
  + SONIA/soNNia for learning sequence-level selection (Q), giving Ppost = Q·Pgen
* **Relevance to autoimmunity**:
  + The methodology can detect sequence sharing not explained by recombination alone, revealing possible convergent immune responses—valuable for **autoimmune biomarker discovery**.
  + Provides a **generalizable framework** for comparing autoimmune vs. healthy repertoires.

**2. Benchmarking Crisis and the Need for Gold Standards**

The paper directly addresses the **lack of quantitative models for repertoire overlap**:

* Shows that using just Pgen underestimates shared sequences.
* Demonstrates how incorporating a selection model Ppost improves predictive power.
* Introduces a **"selection correction factor" q** to account for extra selection pressure not captured by Ppost—used to benchmark model accuracy.

**Gold standard suggestion**:

* Use sharing spectra as a benchmark.
* Fit q across conditions (e.g., CMV+, aging, COVID-19) to benchmark **selection stringency**.

**3. Simulation in the Era of Deep Learning and Multi-Omics**

Although this study uses **probabilistic modeling**, it enables deep learning integration:

* Output of Ppost can serve as prior probabilities for rare sequence generation.
* The model's **sequence-level likelihood** can be fused with omics/disease features in a DL framework.
* Introduces **soNNia**, a deep-learning–based extension of SONIA, trained on >10⁶ sequences.

**Limitation**:  
No direct integration of transcriptomics, HLA type, or spatial data.

**4. Simulation for Personalized Immunology and Precision Medicine**

The study illustrates how **individualized receptor models** can:

* Estimate immune repertoire diversity and convergence (via q).
* Predict **SARS-CoV-2–specific receptors** based on abnormal sharing patterns.
* Identify public antibody or TCR responses, even across individuals with unique HLA backgrounds.

**Towards Digital Twins**:

* Personalized Ppost models could simulate **individual immune landscapes**, enabling:
  + Prediction of autoimmune flares.
  + Tracking immune aging.
  + Designing epitope-based vaccines.

**5. Critical Gaps and Future Directions**

**Identified Gaps**:

* Ppost still requires a correction factor q → Suggests models miss some selection forces (e.g., HLA).
* IgG/IgM analysis based on bulk data → No single-cell lineage resolution.
* No autoimmune cohort modeling yet—authors suggest the method could be extended.

**Future Directions**:

* Incorporate **HLA genotypes** to refine thymic selection modeling.
* Model **longitudinal dynamics** of repertoire change in disease.
* Apply sharing analysis to **autoimmune diseases** (e.g., T1D, SLE, RA).
* Move from identifying shared sequences to **predicting immunogenicity or pathogenicity**.

**Connection to VDJ Individualization (Genome Res. 2021)**

*“VDJ recombination is shaped by both genetics and non-genetic factors.”*

This article supports that insight:

* Pgen (from IGoR) is inferred from individual-specific repertoires.
* Sharing spectra reveal that **even shared sequences follow individual-specific biases**.
* Suggests that **q-factor variation** between individuals or cohorts may be a proxy for **immune personalization**.

**Mathematical Characterization of Private and Public Immune Receptor Sequences**

**Tool/Model Summary Template**

1. **Tool/Model Name:**  
   Mathematical Framework for Public/Private Immune Repertoire Quantification
2. **Reference(s):**  
   Lucas Böttcher, Sascha Wald, Tom Chou.  
   *Mathematical Characterization of Private and Public Immune Receptor Sequences*.  
   *Bulletin of Mathematical Biology* (2023) 85:102.  
   https://doi.org/10.1007/s11538-023-01190-z
3. **Category:**  
   Statistical Modeling, Full Repertoire Characterization, Sampling Theory
4. **Main Purpose/Function:**  
   To mathematically quantify and distinguish **public** (shared across individuals) and **private** (individual-specific) immune receptor sequences, particularly TCR and BCR clones, using probabilistic and information-theoretic measures across individuals and sampling protocols.
5. **Key Features:**
   * Probabilistic, analytical framework
   * Applicable to both TCR and BCR data
   * Supports modeling of sampling effects and subsampling variance
   * Captures richness, overlap (M-publicness), and variance
   * Supports discrete and continuous (e.g., power-law) distributions
   * Includes sampled vs. full repertoire models
   * Simulation-validated
   * Provides formulas for mean, variance, and confidence intervals
6. **Input Requirements:**
   * Clone count or frequency distributions (empirical or synthetic)
   * Sample sizes per individual
   * Optionally: generation probabilities (e.g., from SONIA)
   * Number of individuals (M)
7. **Output:**
   * Richness and overlap metrics (mean, variance)
   * Public vs. private sequence counts
   * Sampled vs. true repertoire characteristics
   * Analytical formulas and confidence intervals
   * Overlap distributions (e.g., Poisson binomial)
8. **Validation/Evaluation:**
   * Compared with simulated synthetic and empirical TCR sequence datasets
   * Validated with SONIA-generated TRB CDR3 sequences
   * Variance and Fano factor computed to estimate reliability
   * Matches observed patterns of clone sharing across individuals
9. **Strengths:**
   * Rigorous, analytical treatment of repertoire sharing
   * Accounts for subsampling bias and variance
   * Provides confidence intervals and expectations
   * Scalable (parallelizable with Numba, supports high-Ω limits)
   * Interpretable mathematical forms
10. **Limitations:**

* Requires prior knowledge or estimation of generation probabilities
* Assumes independence between individuals (no genetic relation)
* No direct modeling of somatic hypermutation dynamics
* Not designed as a user-facing software tool (code available, not GUI)

1. **Applications in Autoimmunity:**

* Not explicitly tested in autoimmune datasets, but the framework is suitable.
* Can help identify whether autoimmune-related clones are public or private across patients.

1. **Notable Citations/Use Cases:**

* This framework builds upon and extends previous works:
  + Elhanati et al., 2018
  + Ruiz Ortega et al., 2023
* Uses SONIA-generated sequences for evaluation

1. **Reviewer’s Comments:**  
   This is a highly rigorous and general-purpose statistical modeling framework. It complements simulation-based methods by offering precise expectations and variances for publicness metrics. Future integration with selection models (e.g., OLGA, IGoR, SONIA) and extension to SHM or class-switching dynamics in BCRs could enhance applicability. Visualization tools or user-friendly interfaces would also broaden impact.
2. **Simulation Level / Output Granularity:**

* Clone-level resolution
* Outputs individual and group-level overlap/richness statistics
* Supports both discrete and continuous trait modeling

1. **Visualization:**

* Not embedded in the tool; figures in the paper include:
  + Richness vs. sampling curves
  + Overlap and publicness (K(M)) metrics
  + Fano factor vs. sample size
* Visualization must be implemented separately (e.g., using Python, R)

**1. Simulation as a Bridge Between Data and Mechanism in Autoimmunity**

**OLGA** serves as a **mechanistically grounded simulation framework**, quantifying the **generation probability** of immune receptor amino acid sequences (e.g., CDR3) based on V(D)J recombination rules. It offers:

* Direct linkage between **genomic models (e.g., IGoR)** and **functional immune receptor diversity**.
* Support for **CDR3 amino acid motifs**, including those implicated in disease.

**Autoimmunity Application**:

* Detects sequences with abnormally high or low generation probabilities.
* Can differentiate **public autoreactive clones** (convergently generated) from rare, disease-associated sequences.
* May assist in predicting which TCRs or BCRs are likely to be present **even without antigen exposure**.

**2. Benchmarking Crisis and the Need for Gold Standards**

OLGA tackles this issue directly by:

* Providing **exact, dynamic programming-based probability estimates**.
* Comparing favorably to Monte Carlo (MC) and exhaustive enumeration:
  + **~700× faster** than MC for amino acid sequences.
  + **99.9% correlation** with true expected frequencies.

**Benchmarking Proposal**:

* Use OLGA’s exact output as a **baseline comparator** for:
  + Experimental frequencies (e.g., from AIRR-seq).
  + Motif enrichment detection.
* Create **synthetic cohorts** with known sequence generation distributions for tool benchmarking.

**3. Simulation in the Era of Deep Learning and Multi-Omics**

While OLGA is **not itself a deep learning tool**, it:

* Enables DL workflows by providing **accurate priors (Pgen, Paa\_gen)**.
* Facilitates **motif frequency estimation** for interpretability validation of models (e.g., CNNs for TCR specificity).
* Forms the **probabilistic layer** in pipelines like SONIA, TAPIR, or TULIP.

Integration Suggestions:

* Pair with multi-omics simulations (e.g., transcriptomics, epitope binding).
* Provide training data for **sequence-to-affinity prediction** networks.

**4. Simulation for Personalized Immunology and Precision Medicine**

OLGA is ideal for **individual-level repertoire modeling** when combined with personalized IGoR models:

* Captures variability in **gene usage**, **deletion patterns**, and **insertion profiles**.
* Computes probabilities for any **CDR3 sequence or motif**, useful in:
  + Vaccine design (which TCRs are likely to exist in the naive pool).
  + **Digital twin simulation** (probabilistic mapping of full repertoire space).

**Precision Autoimmunity Potential**:

* Identifies sequences more likely to emerge spontaneously.
* Helps evaluate which self-reactive BCRs or TCRs are expected **under normal recombination** versus **pathogenic selection**.

**5. Critical Gaps and Future Directions**

**Current Limitations**:

* Does not model **selection processes** (e.g., thymic, peripheral).
* Ignores **somatic hypermutation (SHM)** – BCR evolution post-recombination.
* No dynamic simulation of repertoire development over time.

**Future Development Needs**:

* Add SHM module for BCR affinity maturation simulation.
* Extend to **paired-chain** probability modeling (e.g., α/β or heavy/light).
* Support **non-canonical recombination models** for rare immunodeficiencies or diseases.
* Enable **individual-specific model training** for true personalized simulation.

**Relevance to VDJ Individualization (Genome Res. 2021)**

*Genome Res. 2021 shows VDJ recombination varies even among twins.*

OLGA’s strength lies in its flexibility:

* Can compute generation probabilities using **any IGoR-inferred model**, including those trained on individual-specific datasets.
* Already used to evaluate inter-individual **sequence sharing patterns** and **public repertoire formation**.

**Precision Medicine Insight**: OLGA is foundational for simulating **personal immune repertoires** based on genotype, age, or disease state.

**repgenHMM: a dynamic programming tool to infer the rules of immune receptor generation from sequence data**

**Tool Summary: repgenHMM**

* **Tool/Model Name:**  
  The tool is named **repgenHMM**, short for "repertoire generative Hidden Markov Model."
* **Reference:**  
  It was introduced by Elhanati et al. in 2016 in *Bioinformatics* (DOI: 10.1093/bioinformatics/btw112), with code available at [bitbucket.org/yuvalel/repgenhmm](https://bitbucket.org/yuvalel/repgenhmm).
* **Category:**  
  repgenHMM belongs to the category of **V(D)J recombination simulation** and **probabilistic model inference tools**.
* **Main Purpose/Function:**  
  The tool is designed to infer the probabilistic rules governing immune receptor generation—such as gene usage, deletions, and insertions—based solely on non-productive immune sequences using a Hidden Markov Model (HMM).
* **Key Features:**  
  repgenHMM implements a probabilistic generative model that uses a modified Baum–Welch Expectation-Maximization (EM) algorithm.  
  It supports both TCR α- and β-chains and includes detailed modeling of bi-directional D gene deletions and independent VD and DJ insertions.  
  The tool can generate synthetic sequences, calculate generation probabilities, estimate repertoire entropy, and is implemented in C++ with multithreading support.
* **Input Requirements:**  
  The tool requires a FASTA or plain-text file of non-productive TCR sequences, along with reference germline V, J, and D gene segments—typically from the IMGT database.  
  It also allows optional configuration of maximum insertion/deletion lengths and alignment score thresholds.
* **Output:**  
  repgenHMM outputs learned model parameters such as P(V,J), deletion and insertion distributions.  
  It computes the generation probability for each sequence and can generate synthetic recombination events.  
  Additional outputs include entropy estimates, gene usage frequencies, and statistical correlations.
* **Validation/Evaluation:**  
  The tool has been validated using synthetic datasets with known parameters, where it successfully recovers those values.  
  In comparison to MiXCR, it more accurately reconstructs true insertion and deletion distributions.  
  It has been applied to real TCRα and TCRβ datasets, and its entropy estimations match established biological expectations.
* **Strengths:**  
  repgenHMM provides highly accurate probabilistic modeling of immune receptor generation, integrating over all plausible rearrangement paths rather than relying on a single best alignment.  
  It enables deeper biological interpretation by allowing generation probability estimation, repertoire entropy analysis, and exploration of public clones.  
  Its modeling approach is more biologically realistic than deterministic alignment-based tools.
* **Limitations:**  
  The tool is significantly slower than alignment-based tools such as MiXCR (e.g., ~200 seconds versus ~10 seconds for 50,000 sequences).  
  It does not currently support somatic hypermutation (SHM) or amino acid–level modeling.  
  It requires pre-alignment for V and J genes, and for β-chain modeling, it lacks pre-alignment for D genes, which may reduce accuracy.
* **Applications in Autoimmunity:**  
  While repgenHMM was not originally applied to autoimmune datasets, it is well-suited for modeling baseline repertoire generation in autoimmune vs. healthy individuals.  
  It can be used to detect shifts in rearrangement patterns that may indicate disease-specific immune alterations.  
  The tool also offers a foundation for comparing generative repertoires to those shaped by selection, which is valuable in immunopathology research.
* **Notable Citations and Use Cases:**  
  The tool was introduced in the 2016 publication by Elhanati et al. and is built on earlier work by Murugan et al. (2012).  
  It has been cited in numerous computational immunology studies and is frequently used by the Mora–Walczak group in immune repertoire analysis pipelines.
* **Reviewer’s Commentary:**  
  repgenHMM is considered a foundational tool in computational immune repertoire modeling.  
  It is especially well-suited for non-productive sequence analysis, where selection bias is minimized.  
  Its integration with selection models like SONIA, support for SHM, and amino acid–level generation are recommended as promising future directions.  
  The tool also has strong potential for patient-specific modeling in autoimmunity and other immune-related diseases.
* **Simulation Level / Output Granularity:**  
  repgenHMM simulates the **raw nucleotide-level V(D)J recombination process** and does not produce **AIRR-compliant repertoires** with annotations such as CDR3, isotype, or clonal lineage.  
  It does not simulate downstream processes like clonal expansion, selection, or somatic hypermutation, focusing purely on the generative stage of receptor diversity.
* **Visualization**

repgenHMM does **not include built-in visualization features**. However, its numeric outputs—such as deletion/insertion distributions, gene usage frequencies, generation probabilities, and repertoire entropy—are well-structured and can be **easily visualized using external tools** such as **R** or **Python** libraries (e.g., matplotlib, seaborn, ggplot2)

* **Simulation as a Bridge Between Data and Mechanism in Autoimmunity**

**repgenHMM** reconstructs the **probabilistic rules of V(D)J recombination** directly from high-throughput AIRR-seq data using a **Hidden Markov Model (HMM)**. It provides:

* Mechanistic insight into recombination: gene usage, deletion, and insertion statistics.
* The ability to **generate synthetic repertoires** that mimic biological realism.

**Autoimmune connection**: The tool captures sequence generation probabilities, allowing researchers to assess whether **disease-associated clonotypes** are statistically likely or selected—essential for identifying **autoreactive expansions**.

* **Benchmarking Crisis and the Need for Gold Standards**

repgenHMM provides:

* **Exact likelihood-based models** trained on real data.
* A **synthetic sequence generator** that can serve as a **ground-truth simulator**.
* Validation against synthetic data (with known parameters) confirms **high recovery accuracy**.

**Implication**: repgenHMM is an ideal foundation for **benchmarking other annotation or generative tools** (like MiXCR, IGoR, OLGA) using **ground-truth-controlled synthetic datasets**.

* **Simulation in the Era of Deep Learning and Multi-Omics**

While repgenHMM is **not DL-based**, it offers:

* A **model-based probability landscape** (Pgen) for any sequence.
* Ground-truth training data for DL models, e.g., immune sequence classification or specificity prediction.

**Integration potential**:

* Feed repgenHMM-generated repertoires into **DL models** (e.g., BERT-based TCR models).
* Can be paired with **SHM models** (external) to simulate BCR evolution for downstream omics integration.
* **Simulation for Personalized Immunology and Precision Medicine**

repgenHMM supports:

* **Individual-specific model inference** from non-productive repertoires.
* Reconstruction of person-specific Pgen distributions—a foundation for **digital immune twins**.

**Clinical potential**:

* Assess if autoreactive or shared sequences are **likely vs. selected**.
* Track deviations in VDJ statistics for **autoimmune disease monitoring**.
* Use synthetic outputs to simulate **baseline repertoires** across genetic backgrounds.
* **Critical Gaps and Future Directions**

**Gaps**:

* No explicit support for **SHM**, **selection modeling**, or **paired chains**.
* Sequence outputs are **nucleotide-based only** (no built-in amino acid simulation).
* Lacks modular integration with affinity or structure-based tools.

**Future Enhancements**:

* Incorporate **context-dependent SHM** modeling.
* Extend to **paired-chain inference** and **structural prediction compatibility**.
* Add modules for **multi-chain receptor generation** and **temporal immune dynamics**.

**Relevance to Individualized VDJ Recombination (Genome Res. 2021)**

*"VDJ recombination is influenced by both genetic and environmental factors."*

repgenHMM aligns with this vision:

* It infers **individual-specific V, D, J usage**, insertions, and deletion profiles.
* Can be trained on **any individual's nonproductive sequences**, modeling personalized immune recombination rules.
* Ideal for capturing immune individuality—even in monozygotic twins, as the authors suggest.

**Mathematical Characterization of Private and Public Immune Receptor Sequences**

**Tool/Model Summary Template – OLGA**

1. **Tool/Model Name:**  
   **OLGA** (Optimized Likelihood estimate of immunoGlobulin Amino-acid sequences)
2. **Reference(s):**  
   Sethna Z, Elhanati Y, Callan CJ Jr, Walczak AM, Mora T.  
   *OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs*.  
   *Bioinformatics*, 2019, 35(17):2974–2981.  
   DOI: 10.1093/bioinformatics/btz035  
   GitHub: <https://github.com/zsethna/OLGA>
3. **Category:**
   * V(D)J Recombination Model
   * Probability Estimator
   * Motif Analysis
   * Benchmarking Tool
4. **Main Purpose/Function:**  
   OLGA calculates the **probability of generation (Pgen)** for **amino acid sequences or motifs** of BCR or TCR via V(D)J recombination, overcoming computational limitations posed by codon degeneracy. It enables evaluation of how likely a specific receptor sequence is to occur naturally.
5. **Key Features:**
   * **Probabilistic and exact** computation (not MC-based)
   * Based on **dynamic programming**
   * Computes **Pgen** for full sequences or motifs (with or without V/J constraints)
   * Handles **TCRα, TCRβ, BCR heavy chains** (human and mouse)
   * Can use **motif patterns** (e.g., regex-like)
   * Fast: ~47 sequences/sec/CPU
   * Compatible with models trained by IGoR
6. **Input Requirements:**
   * CDR3 amino acid sequences or motif patterns
   * A trained generative model (e.g., from IGoR)
   * Optional: restricted V/J gene usage
7. **Output:**
   * Generation probability of input sequences or motifs
   * Summary statistics for epitope-specific sequence sets
   * Entropy and distribution of probabilities across loci
   * Heatmaps and frequency plots (via external tools)
8. **Validation/Evaluation:**
   * Validated against Monte Carlo simulations: excellent correlation (KL divergence ≈ 4.8×10⁻⁷ bits)
   * Compared against IGoR and MC: OLGA is **~700× faster**
   * Tested across loci (TRA, TRB, IGH) and species (human, mouse)
   * Cross-validated with real data from large repertoires (Emerson et al.)
9. **Strengths:**
   * **Highly accurate and efficient** for computing Pgen of amino acid sequences
   * **Works on motifs**, not just exact sequences
   * Enables **epitope-level repertoire analysis**
   * Avoids bias of MC estimation or sequence enumeration
   * Facilitates **vaccine design** and **cross-species predictions**
10. **Limitations:**

* Does **not model SHM** (somatic hypermutation)
* Requires **pre-trained model** (e.g., from IGoR)
* Chain pairing (TCRα–β) handled only independently
* CLI-based; no GUI
* Limited to known VDJ loci unless extended manually

1. **Applications in Autoimmunity:**

* Not designed specifically for autoimmunity but **suitable**
* Can identify **public vs. private clonotypes**
* Used to evaluate generation probabilities of TCRs found in **autoimmune diseases** (e.g., ankylosing spondylitis, MS, type 1 diabetes)
* Aids in distinguishing disease-driven expansions from randomly generated sequences

1. **Notable Citations/Use Cases:**

* Emerson et al., 2017: CMV-associated TCRs
* Dash et al., 2017: Epitope-specific TCR motif analysis
* Pogorelyy et al., 2018: TCR sharing and convergent recombination
* Glanville et al., 2017: Vaccine response TCR clustering

1. **Reviewer’s Comments:**  
   OLGA is a gold-standard tool for efficient and exact estimation of generation probabilities at the amino acid level. Its performance and flexibility in motif handling make it indispensable in immune repertoire analysis. Incorporating SHM modeling and GUI-based usability would further broaden its impact.
2. **Simulation Level / Output Granularity:**

* **Amino acid level** (CDR3)
* Motif-based querying
* Aggregated statistics for epitope-level groups
* Works across multiple V(D)J loci

1. **Visualization:**

* Not built-in
* External tools (e.g., Python, R) can plot:
  + Pgen distributions
  + Entropy plots
  + Heatmaps for motif frequency
  + Scatter plots for validation (e.g., MC vs. OLGA Pgen)

**1. Simulation as a Bridge Between Data and Mechanism in Autoimmunity**

This article doesn’t present a simulation tool per se, but proposes a **mathematical modeling framework** that:

* Quantifies **publicness and privateness** in TCR/BCR repertoires.
* Offers analytical predictions for clonal overlap and richness across individuals.
* Enables comparison between **theory and empirical/simulated data**, including AIRR-seq.

**Bridge to Mechanism**:  
It supports **mechanistic insight** by modeling how clone generation probabilities affect public/private clonotypes. It aids clinical translation by predicting:

* Shared vs. private TCR/BCR patterns in disease cohorts.
* Sampling effects on observed diversity.

**2. Benchmarking Crisis and the Need for Gold Standards**

The authors highlight that:

* Current definitions of “public” vs. “private” clones are **inconsistent**.
* Their analytical metrics (e.g., **M-overlap**, **expected richness**, **Fano factors**) offer **standardizable benchmarks**.

**Proposal for Benchmarking**:

* Use **mathematically defined overlap statistics** as **ground truth** for validating sampling, diversity estimation, and clone sharing tools.
* Compare tools under controlled simulations (as they did with SONIA-generated data and synthetic data).

**3. Simulation in the Era of Deep Learning and Multi-Omics**

Although not a DL-based paper, it provides:

* **Theoretical foundations** for evaluating DL models trained on repertoire data.
* Metrics like clone richness, overlap, and variance to assess DL-based generative models (e.g., Immune2vec, TCR-BERT).

**Use Case**:

* Supports **data augmentation** and **validation** pipelines in DL tools by offering expected statistical distributions of TCR/BCR repertoires.

**4. Simulation for Personalized Immunology and Precision Medicine**

The paper is well-positioned for **personalized immunology**:

* Defines **individual-specific clone probabilities** pi(m)p\_i^{(m)}pi(m)​ and models how they affect observed repertoires.
* Uses **Bayesian inference** (Eq. 42–44) to reconstruct true diversity from small blood samples.

**Digital Twin Potential**:

* Enables prediction of public vs. private clone structure in **individuals** or **disease subgroups**.
* Can model how sampling bias might hide or exaggerate **autoimmune clone expansions**.

**5. Critical Gaps and Future Directions**

The authors suggest several directions:

* Need for more **realistic clone generation models** incorporating selection, SHM, and antigen exposure.
* Extend from single-chain modeling to **paired-chain** and **longitudinal tracking**.
* Account for **coarse-graining and information loss** in AIRR-seq and spectratyping.

**Call to Action**:

* Use the proposed framework to **standardize immune repertoire metrics**.
* Integrate with empirical data from **autoimmune diseases**, **vaccine studies**, and **T-cell engineering**.

**Relevance to VDJ Recombination Individuality**

*Genome Res. 2021 showed that VDJ rules vary even among monozygotic twins.*

**Inference of B cell clonal families using heavy/ light chain pairing information**

**Tool/Model Summary Template**

**Tool/Model Summary Template**

1. **Tool/Model Name:**  
   **Paired Clonal Family Inference** (implemented within the *partis* framework)
2. **Reference(s):**  
   Ralph DK, Matsen FA IV. *Inference of B cell clonal families using heavy/light chain pairing information*. PLOS Computational Biology. 2022; 18(11): e1010723.  
   <https://doi.org/10.1371/journal.pcbi.1010723>
3. **Category:**  
   Clonal Inference Tool; Repertoire Partitioning; Benchmarking Tool; Phylogenetic Support Model
4. **Main Purpose/Function:**  
   To accurately infer B cell clonal families from AIRR-seq data by incorporating both heavy and light chain pairing information, improving clonal assignment and overcoming ambiguity due to limited light chain diversity.
5. **Key Features:**

* Probabilistic, hierarchical model
* Handles both **paired-chain** and **unpaired** datasets
* Compatible with **partis** framework
* Modular and scalable
* Supports benchmarking against multiple inference strategies
* Offers novel subclustering and ambiguity resolution algorithms
* Improves accuracy in high-throughput droplet-based single-cell data

1. **Input Requirements:**

* BCR sequences (heavy and/or light chains)
* AIRR-formatted data (FASTA/TSV with annotations)
* Optional: single-cell barcodes or pairing metadata

1. **Output:**

* Clonal family assignments
* Subcluster annotations
* Pairing confidence metrics
* Precision-recall evaluations (optional)
* JSON and TSV summary statistics for downstream analysis

1. **Validation/Evaluation:**

* Validated using both simulated and real experimental datasets
* Benchmarked against SCOPer, enclone, MobiLLe
* Evaluated using F1 score, overmerging and oversplitting rates
* Demonstrated improvements in disambiguating multi-cell droplets

1. **Strengths:**

* Effectively leverages light/heavy chain pairing to improve accuracy
* Supports integration of single-cell and bulk data
* Provides interpretable and explainable metrics
* Scalable to large repertoire datasets
* Incorporates biological priors (e.g., gene usage, SHM distribution)

1. **Limitations:**

* Does not model SHM explicitly (only indirectly through distance metrics)
* No antigen binding or structure-level prediction
* Performance may be limited by poor-quality pairing information
* Does not handle isotype switching or longitudinal dynamics

1. **Applications in Autoimmunity:**

* While not applied directly to autoimmune disease in this study, the model is suitable for tracing clonally expanded autoreactive B cells in diseases like SLE, RA, MS.
* Particularly useful for studies requiring accurate clone tracking across compartments or timepoints.

1. **Notable Citations/Use Cases:**

* This study itself is the primary citation; model integrated into the *partis* software suite.
* Useful for groups studying single-cell B cell evolution, vaccine response, and clonal expansions in infection.

1. **Reviewer’s Comments:**

* This method addresses a critical bottleneck in BCR repertoire analysis: pairing-based ambiguity.
* Future integration with antigen-binding predictors (e.g., paratope modeling) or SHM-aware models could extend its utility.
* An important step toward higher-fidelity digital immune repertoire analysis in both health and disease contexts.

1. **Simulation Level / Output Granularity:**

* High resolution: clone-level and subclone-level output
* Works at nucleotide-level granularity, with amino acid-level cluster similarity options

1. **Visualization:**

* No native GUI or plots, but outputs are compatible with visualization libraries (e.g., ggplot2, matplotlib)
* Clone trees and cluster heatmaps can be derived from outputs using external tools

**1. Simulation as a Bridge Between Data and Mechanism in Autoimmunity**

This paper doesn’t introduce a simulator in the classic sense (like IGoR or immuneSIM), but it **enables better interpretation of repertoire data** by improving **clonal family inference using pairing information** (i.e., linking heavy and light chain data). Such accurate partitioning is essential for modeling mechanisms such as clonal selection, expansion, and SHM in autoimmunity. By resolving ambiguities in clonal assignments — especially in light chains where diversity is limited — the method facilitates more **mechanistically faithful interpretations of AIRR-seq datasets**, which is crucial for understanding autoreactive B cell development.

**2. Benchmarking Crisis and the Need for Gold Standards**

The authors **emphasize rigorous benchmarking**, using both simulated and real datasets to test their algorithm. They propose **F1 score-based metrics** to quantify overmerging and oversplitting — two major issues in clonal inference — and test against state-of-the-art tools like SCOPer, enclone, and MobiLLe. This provides a **valuable framework for community benchmarking**, which is currently lacking in repertoire analysis. Their simulation-matching approach and transparent use of GitHub (<https://github.com/psathyrella/partis>) are steps toward **standardized, reproducible benchmarking pipelines**.

**3. Simulation in the Era of Deep Learning and Multi-Omics**

While the tool itself does not use deep learning, the **paired clustering outputs** it generates are foundational for downstream **ML-based repertoire modeling** (e.g., learning patterns of SHM or predicting antigen binding). It also opens doors for **multi-omic integration** — such as matching repertoire data with single-cell transcriptomes — by resolving clone identity more accurately, especially important when BCRs are linked with phenotypic or transcriptional traits in multi-modal datasets.

**4. Simulation for Personalized Immunology and Precision Medicine**

The ability to **disambiguate multiple cells per droplet** and **approximately pair bulk and single-cell samples** means that the tool can scale to larger, patient-specific datasets. This supports efforts to develop **digital immune twins**, where both heavy and light chains need to be accurately paired to model a patient’s antibody repertoire. Furthermore, the **subcluster annotation improvements** make it more robust to individual-specific phylogenetic structures, which is critical for modeling autoimmune flares, vaccine responses, or therapy design on a personalized level.

**5. Critical Gaps and Future Directions: What the Field Needs Next**

The authors call for future developments in:

* **Handling allelic inclusion and multi-chain expression** (currently not modeled),
* **Integrating paired clustering directly into the inference process**, and
* **Evaluating biological pairing correlations** — a crucial gap in understanding BCR evolution.

They also highlight that **paired clustering accuracy becomes critical** as dataset sizes grow and **VDJ collisions increase**, especially for light chains. Hence, their work is a call to **build more intelligent, scalable, and integrative simulation+inference platforms**.

This paper **does not simulate repertoires per se**, but it addresses a **critical bottleneck in simulation utility**: the ability to accurately infer clonal relationships from data, particularly in **paired AIRR-seq experiments**. Its contributions in resolving pairing ambiguity and benchmarking methods make it a foundational piece for any simulation-driven or AI-enhanced immunological modeling — especially as the field moves toward **precision autoimmunity** and **digital twin frameworks**.

**sumrep: A Summary Statistic Framework for Immune Receptor Repertoire Comparison and Model Validation**

**Tool/Model Summary Template**

1. **Tool/Model Name:**  
   *Statistical Repertoire Sharing Model (Pgen + SONIA / soNNia)*
2. **Reference(s):**  
   Ruiz Ortega M, Spisak N, Mora T, Walczak AM (2023). *Modeling and predicting the overlap of B- and T-cell receptor repertoires in healthy and SARS-CoV-2 infected individuals.* PLoS Genet 19(2): e1010652.  
   <https://doi.org/10.1371/journal.pgen.1010652>
3. **Category:**  
   Repertoire Sharing Prediction / Statistical Modeling / Selection Modeling
4. **Main Purpose/Function:**  
   To quantitatively model and predict the sharing (publicness) of immune receptor sequences across individuals using generative and selection models. The method can also identify antigen-specific (e.g., SARS-CoV-2) receptors by detecting overshared sequences beyond chance expectations.
5. **Key Features:**
   * Probabilistic and interpretable
   * Trained on non-productive (Pgen) and productive (Ppost) repertoires
   * Selection modeling using SONIA (linear) or soNNia (neural networks)
   * Can detect overshared, condition-associated clonotypes
   * Includes correction factors (q) for convergent selection
   * Validated across BCR and TCR data
6. **Input Requirements:**
   * AIRR-seq data from productive and non-productive repertoires
   * Annotated V(D)J sequences (e.g., CDR3, V/J genes)
   * Optional: metadata (e.g., age, CMV status, infection status)
7. **Output:**
   * Predicted sharing spectrum
   * Posterior probabilities of sequence generation (Pgen, Ppost)
   * Identified overshared clonotypes
   * Diagnostic likelihood scores
   * Network visualizations of shared sequences
8. **Validation/Evaluation:**
   * Validated using datasets from healthy and SARS-CoV-2–infected individuals
   * Comparison against known SARS-CoV-2–specific antibody/TCR databases
   * ROC curves and classification tests used for diagnostic potential
9. **Strengths:**
   * High predictive accuracy for BCR sharing
   * Integration of generative and selection pressures
   * Capable of identifying public antigen-specific sequences
   * Adaptable to different disease contexts and age groups
10. **Limitations:**

* TCR modeling requires correction factor (q), indicating missing selection signals
* Doesn’t yet fully model HLA-type-specific selection
* Limited validation on longitudinal or single-cell data
* Diagnostic predictions for TCRs less robust than BCRs

1. **Applications in Autoimmunity:**  
   While the paper focuses on SARS-CoV-2, the framework is directly applicable to autoimmune diseases for identifying public autoantigen-specific receptors and understanding convergent selection due to autoreactivity.
2. **Notable Citations/Use Cases:**

* This paper itself (Ruiz Ortega et al., 2023)
* Extension of foundational work by Elhanati et al., 2018 (Pgen) and Marcou et al., 2018 (SONIA)

1. **Reviewer’s Comments:**  
   This model represents a powerful combination of theory-driven and data-driven approaches. It excels in B cell repertoire prediction, and the use of statistical “correction” factors in T cells points toward a roadmap for future refinement using HLA and longitudinal data. Its diagnostic capability is particularly compelling, with potential applications beyond infectious disease.
2. **Simulation Level / Output Granularity:**

* Clonotype-level predictions
* Cohort-level sharing spectra
* Sequence-level selection probability scores (Ppost)

1. **Visualization:**

* Sharing spectra plots
* Scatter plots of empirical vs. predicted frequency
* Network visualizations of clonotype similarity clusters
* ROC curves for diagnostics

**1. Simulation as a Bridge Between Data and Mechanism in Autoimmunity**

Although not a simulator in the classical sense, the Ruiz Ortega model plays a vital bridging role between immune receptor generation mechanisms and observable immune data. By combining Pgen (generation probability) and Ppost (post-selection probability) via tools like SONIA and soNNia, the model helps disentangle stochastic generation from antigen-driven selection. This layered approach is especially powerful in autoimmunity, where public clonotypes may arise from both convergent recombination and shared self-antigen exposure. The framework enables hypothesis testing at the cohort level—for example, determining whether shared clonotypes among autoimmune patients exceed expectations from neutral generation, pointing toward disease-relevant convergence.

**2. Benchmarking Crisis and the Need for Gold Standards**

This work exemplifies a robust benchmarking approach by validating predictions against known SARS-CoV-2–specific clonotypes and performing ROC analyses for diagnostic classification. It also introduces interpretable metrics like the “correction factor” qqq, which quantifies selection intensity beyond generative expectation. Such metrics provide a reproducible and generalizable way to assess repertoire overlap, filling a gap in standardized evaluation. While most benchmarks focus on sequence alignment or generation models alone, this framework integrates multiple probabilistic layers—offering a pathway to benchmark future simulation-based disease classifiers or publicness detectors.

**3. Simulation in the Era of Deep Learning and Multi-Omics**

The model’s use of **soNNia**, a neural network–based selection model, illustrates how classical generative modeling can interface with modern deep learning approaches. Moreover, its structure is modular and data-agnostic, making it amenable to integration with multi-omics profiles such as HLA genotypes, transcriptomic clusters, or phenotypic markers. This allows for simulation-guided statistical learning rather than purely black-box predictions, an essential quality when interpretability is required—for example, when validating autoreactive or vaccine-associated clonotypes in autoimmune or post-infectious settings.

**4. Simulation for Personalized Immunology and Precision Medicine**

The model is designed to train on **individual-specific** repertoire data (productive and non-productive), enabling personalized inferences about generation likelihood, selection pressure, and publicness. This makes it highly suitable for **digital twin frameworks**, where understanding whether a patient’s autoreactive clone is expected or selected can shape diagnostics and treatment design. The ability to distinguish between statistical rarity and immunological relevance—especially for highly shared public clonotypes—has clear implications for detecting emerging autoimmune flares or monitoring therapeutic response.

**5. Critical Gaps and Future Directions**

Despite its innovation, the framework has limitations—particularly for TCR modeling, which requires correction factors like qqq to account for unseen HLA- or context-dependent selection. Moreover, it lacks native support for paired-chain modeling and longitudinal dynamics, both of which are increasingly important in autoimmune research. Future directions should focus on expanding its utility across TCR modalities, integrating it with single-cell and temporal datasets, and embedding it in broader simulation pipelines. Nevertheless, its transparent modeling, selection-aware probabilistic structure, and validation capacity make it a valuable benchmark-setting tool for repertoire-based autoimmunity research.

**Comparative Review of Immune Repertoire Simulation and Modeling Tools in Autoimmunity**

**Integrating Paired Clonal Inference into the Repertoire Simulation Landscape**  
While repertoire simulation tools such as IGoR, OLGA, repgenHMM, immuneSIM, and AIRRSHIP provide valuable frameworks for modeling the generation and evolution of B and T cell receptors, they typically focus on sequence-level probabilistic modeling, synthetic repertoire generation, or somatic hypermutation dynamics. In contrast, the paired clonal inference approach developed by Ralph and Matsen (2022) addresses a distinct but critical aspect of immune repertoire analysis: the accurate partitioning of observed sequences into clonal families, particularly in the context of heavy and light chain pairing. This method does not generate new sequences but rather enhances the biological interpretation of high-throughput AIRR-seq data by resolving pairing ambiguities and improving the granularity of inferred clonal structures. As such, it serves as a foundational post-processing module that strengthens the downstream utility of simulated or experimental data in both health and disease studies.

**Benchmarking Accuracy and Applications in Autoimmunity**  
A distinguishing feature of this tool lies in its rigorous benchmarking framework, which evaluates clonal inference performance using simulated and empirical datasets. The model demonstrates improved precision in avoiding common artifacts such as overmerging and oversplitting—issues that can significantly compromise interpretations in autoimmune disease studies where clonal expansion is a hallmark. While simulation-based tools like immuneSIM or AIRRSHIP provide ground-truth repertoires for method testing, the paired inference method introduces robust evaluation metrics (e.g., F1 score) specifically tailored to clonal family resolution. Furthermore, its ability to integrate both droplet-based single-cell data and bulk sequencing enhances its applicability in longitudinal and personalized immune monitoring—particularly relevant for profiling autoreactive B cell responses.

**Enhancing Deep Learning and Precision Immunology Pipelines**  
Although not built on deep learning itself, the paired inference framework complements AI-driven models such as TAPIR and TULIP by providing high-confidence clonal annotations that can serve as curated input for training and evaluation. As deep learning approaches increasingly rely on accurate and interpretable data for modeling immune specificity or repertoire dynamics, tools that enhance data quality—especially in the context of paired-chain information—are becoming indispensable. Additionally, the method’s modular implementation and compatibility with existing single-cell platforms position it well within the emerging paradigm of digital twin immunology. In such frameworks, individualized modeling of clonal evolution is essential for predicting autoimmune flare dynamics, therapeutic responsiveness, and immune memory. Thus, the integration of this inference tool alongside simulation engines and predictive models offers a more complete and biologically grounded computational immunology pipeline.

**Summary and Outlook**

In summary, the integration of paired clonal inference methodologies with established immune repertoire simulation and prediction tools represents a critical advancement toward a comprehensive and biologically grounded framework for adaptive immune analysis. While simulators such as IGoR, immuneSIM, and AIRRSHIP facilitate the generation and benchmarking of repertoires under controlled parameters, paired inference tools enhance downstream interpretability—particularly in the context of single-cell data and autoimmune research. As computational immunology moves increasingly toward personalized, multi-omic, and clinically relevant applications, methods that provide accurate clonal architecture will be essential for transforming simulated repertoires into actionable biomedical insights. Future directions should emphasize the convergence of generation models, clonal inference, selection dynamics, and structural prediction into unified, interoperable platforms capable of supporting digital immune twin development and enabling precision medicine in autoimmunity.

Table 1. Comparative overview of key immune repertoire simulation and modeling tools with relevance to autoimmunity, benchmarking, deep learning integration, and personalized immunology.

This table summarizes nine widely used or recently developed tools and models: IGoR, OLGA, immuneSIM, AIRRSHIP, repgenHMM, TAPIR, TULIP, Ruiz Ortega et al. (2023), and Böttcher et al. (2023). Each tool is evaluated across thematic axes including its category (e.g., simulator, inference engine, or deep learning model), simulation capabilities, support for individual-specific modeling, compatibility with deep learning or multi-omics integration, utility for benchmarking, relevance to autoimmune disease research, and current limitations. The comparison highlights how different tools serve complementary purposes—from mechanistic generation and annotation to functional prediction and cohort-level statistical modeling—underscoring the need for integrated platforms in precision immunology.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Tool/Model** | **Category** | **Simulation** | **Personalization** | **DL/Multi-Omics Integration** | **Benchmarking Utility** | **Autoimmunity Relevance** | **Limitations** |
| **IGoR** | V(D)J inference & simulation | Yes (V(D)J + SHM) | Yes (train on individual data) | Enabler for DL (provides training data) | Yes (generates truth sets) | Yes (sequence-based auto-reactivity) | No antigen modeling; no structure |
| **OLGA** | Pgen calculator | No (Pgen only) | Yes (via IGoR) | Supports DL training | Yes (exact Pgen calculator) | Yes (Pgen-based risk estimation) | No SHM or affinity model |
| **immuneSIM** | Full repertoire simulator | Yes (BCR/TCR, tunable) | Yes (parameter control) | DL-compatible (motif implant) | Yes (ground-truth annotated data) | Yes (aberrant repertoire simulation) | No binding or antigen info |
| **AIRRSHIP** | BCR repertoire simulator | Yes (BCR heavy chain) | Limited (global stats) | Limited | Yes (benchmark annotation tools) | Yes (simulate autoreactive SHM patterns) | No TCR support; no antigen modeling |
| **repgenHMM** | Recombination model inference | Yes (sequence generator) | Yes (trained on personal data) | Enabler (DL can use outputs) | Yes (train/test with synthetic) | Yes (generation vs. selection) | No SHM; no structure; nucleotide only |
| **TAPIR** | TCRâ€“epitope prediction (DL-based) | No (predictive model) | No | DL model (CNN) | Yes (functionality benchmarks) | Yes (predicts activation of auto-clones) | No structure; no paired-chain modeling |
| **TULIP** | TCRâ€“epitope generation & prediction (Transformer) | Yes (generative + predictive) | No | Transformer model | Yes (unseen epitope generalization) | Yes (simulates disease-related TCRs) | No affinity prediction; training limited |
| **Ruiz Ortega et al. 2023** | Public clone modeling (Pgen + Ppost) | No (sampling model) | Yes (personal Ppost models) | DL-friendly (soNNia extension) | Yes (selection correction factor q) | Yes (used to infer abnormal publicness) | Requires Pgen model input; no structure |
| **Böttcher et al. 2023 (Mathematical Model)** | Analytical model of public/private clones | No (analytical, but supports synthetic) | Yes (individual clone probability modeling) | Supports DL evaluation (publicness metrics) | Yes (analytical benchmarks: M-publicness, richness) | Yes (quantifies clone sharing in disease) | No dynamics; no sequence-level selection |