BMS PROJECT

Computation Tool to Estimate Charge Variants of a Biopharmaceutical Protein

RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

Prresented By:
Glory E
Mushira S
Pranav S

AGENDA

- **Charge variants in biopharmaceutical proteins impact product stability and efficacy.**
- **Need a computational tool to estimate overall charge distributions efficiently and accurately.**

Phase 1

Phase 2

PROBLEM STATEMENT

CORE ALGORITHM DESIGNS

OFFLINE IMPLEMENTATION

PHASE 2:

VISUALIZATION | CROSS COMPARISON | OPTIMIZATION | INTEGRATION

Convolution

Key Idea

- Each copy = Independent discrete random variable over integer charges.
- The total-charge distribution is the discrete convolution of all copy PMFs. (convolve)

Multiple Copies at a Site

- If a site has n identical copies, its site-level PMF is the n-fold self-convolution of the base PMF.
- We compute this with exponentiation-by-squaring a.k.a. binary powering. (power_pmf)
- This reduces the number of convolutions from O(n) to O(log n).

Complexity & Performance

- Let K be the average support size (e.g., ≤ 11 bins for −5…+5).
- Each convolution costs $O(K^2)$; each power_pmf costs $O(K^2 \log n_j)$.
- Memory footprint is O(K) for the PMFs.

$$\text{Time} \approx O\Big(K^2 \sum_{j=1}^{S} \log n_j \; + \; K^2 \, S\Big)$$

Fast Fourier Transform Convolution

Key Idea

- Uses the Convolution Theorem: convolution in time domain = multiplication in frequency domain.
- Converts each PMF (charge distribution) to frequency space using the Fast Fourier Transform (FFT).
- Performs pointwise multiplication of transformed distributions.
- Applies Inverse FFT (IFFT) to return to the probability domain.
- Significantly speeds up computation while preserving accuracy.

Multiple Copies at a Site

- For sites with n identical copies, FFT efficiently handles self-convolution of the PMF.
- Instead of n repeated convolutions, the transformation is applied once, and multiplication in frequency space combines all copies.

Complexity & Performance

- Each FFT and IFFT costs O(N log N), where N = number of bins (charge states).
- Much faster than standard convolution ($O(N^2)$ per step).

Monte Carlo Simulation Approach

Key Idea

- Based on random sampling to estimate the overall charge distribution.
- Relies on the Law of Large Numbers — with enough samples, the estimated probabilities approach the true distribution.
- Each trial randomly picks charge states from each PTM site's PMF and sums them to get a total charge.
- Repeated thousands of times to approximate the probability distribution.

Multiple Copies at a Site

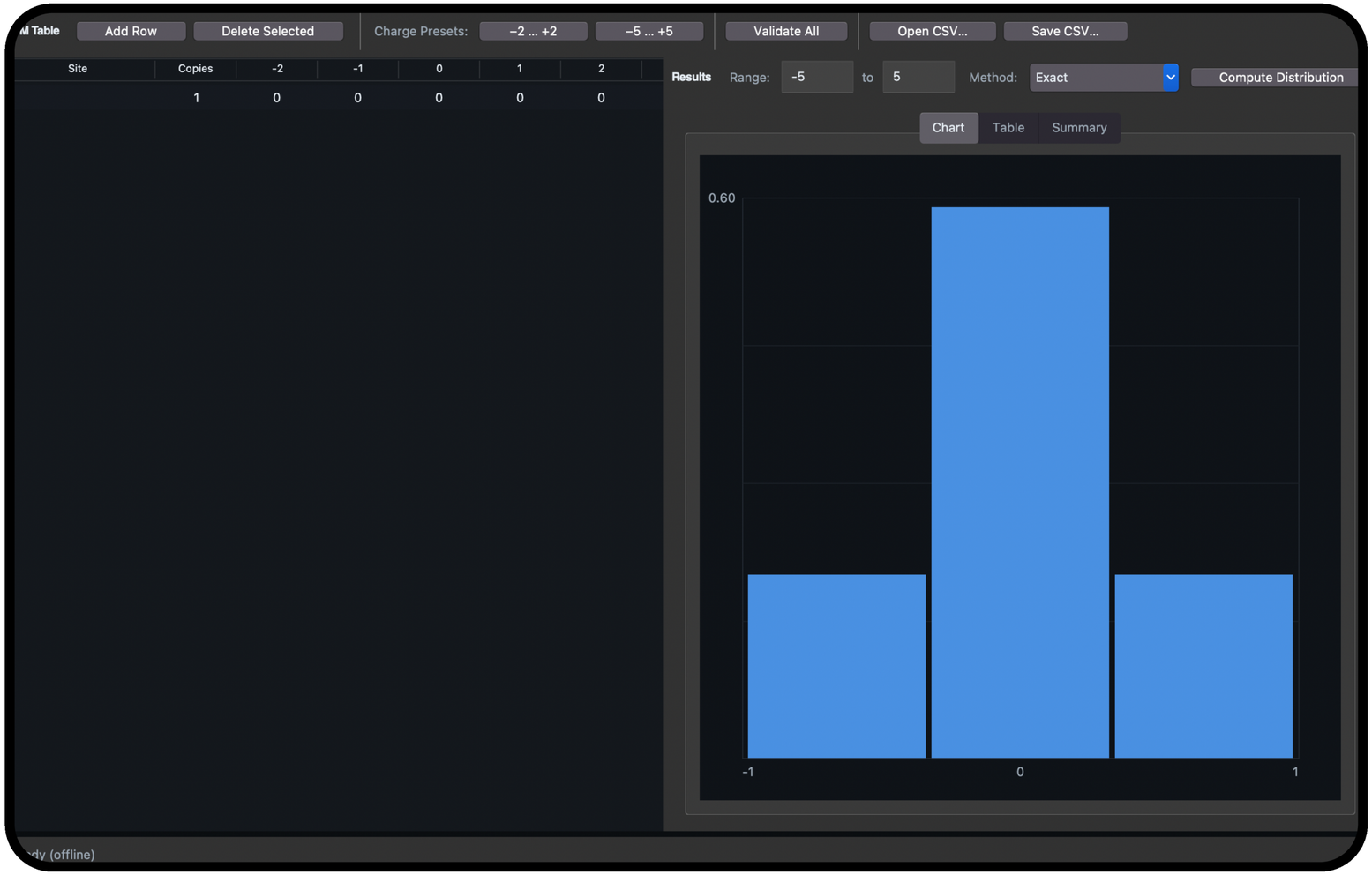
- For each PTM site with n copies, charges are independently sampled n times from its distribution.
- The total charge for that site is the sum of sampled values.
- This is repeated for all sites across multiple random trials to build an empirical total-charge distribution.

Complexity & Performance

- Time Complexity: $O(N \cdot K)$, where N = number of samples and K = number of charge states.
- It becomes slower when N increases
- Fails when N is small — estimates become noisy, unstable, or biased.

OFFLINE

- **User Interface (Frontend) –** Intuitive dashboard for PTM data input, charge presets, validation, and visualization.
- **Computation Engine (Core Logic) –** Implements combinatorial and probabilistic algorithms to estimate overall charge distributions efficiently.
- **Data Handling (Storage & Export) –** Manages user inputs and results locally with Excel/CSV export — no external database required.
- **Offline Deployment (Standalone Tool) –** Packaged with PyInstaller for full offline operation; runs without internet or additional runtime installations.



PHASE II:

ALGORITHM OPTIMIZATION | INTEGRATING

Future development will focus on optimizing the core computation engine using parallel processing and approximation algorithms to handle larger PTM datasets with minimal runtime. Additionally, the tool can be integrated with existing laboratory data systems or analytical pipelines to enable seamless data import, automated analysis, and report generation — bridging in-silico modeling with real-world experimental workflows.

VISUALIZATION | INTEGRATION

In future iterations, the tool can include an interactive algorithm visualization module to display how charge combinations evolve as PTMs increase. Users could watch the distribution build dynamically through animated probability plots or heatmaps, helping scientists intuitively understand how individual site variations contribute to overall protein charge behavior.

EFFICIENCY

The algorithm efficiently estimates overall protein charge distributions by using optimized convolution-based methods instead of brute-force enumeration. This reduces the computational cost from exponential to near-quadratic or log-linear time, making it feasible to handle large PTM datasets. Overall, it achieves a strong balance between speed, accuracy, and low memory usage, ensuring scalable and reliable offline performance.

Thank You

Any Questions ?