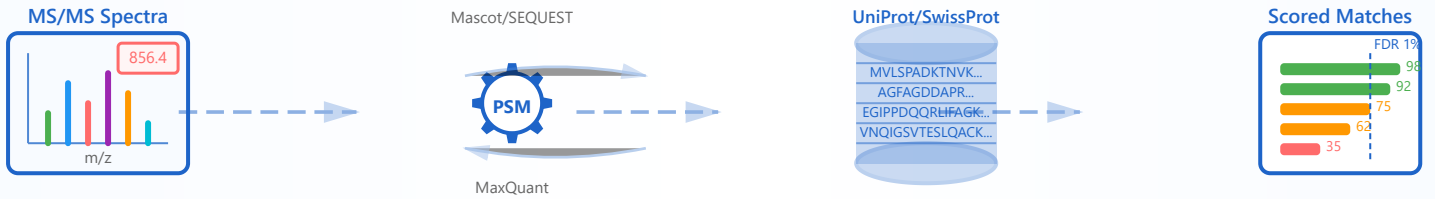
Database Searching



Experimental Spectra
Raw MS/MS data

Search Engine
Scoring algorithm

Protein Database
Reference sequences

Scored Matches
PSM ranking

Search Engines

- Mascot, SEQUEST, X!Tandem
- MaxQuant, Proteome Discoverer
- Each with unique algorithms

Parameter Optimization

- Mass tolerance settings
- Enzyme specificity
- Missed cleavages allowed

Decoy Databases

- Reversed/shuffled sequences
- Estimate false positives
- Quality control

Modifications

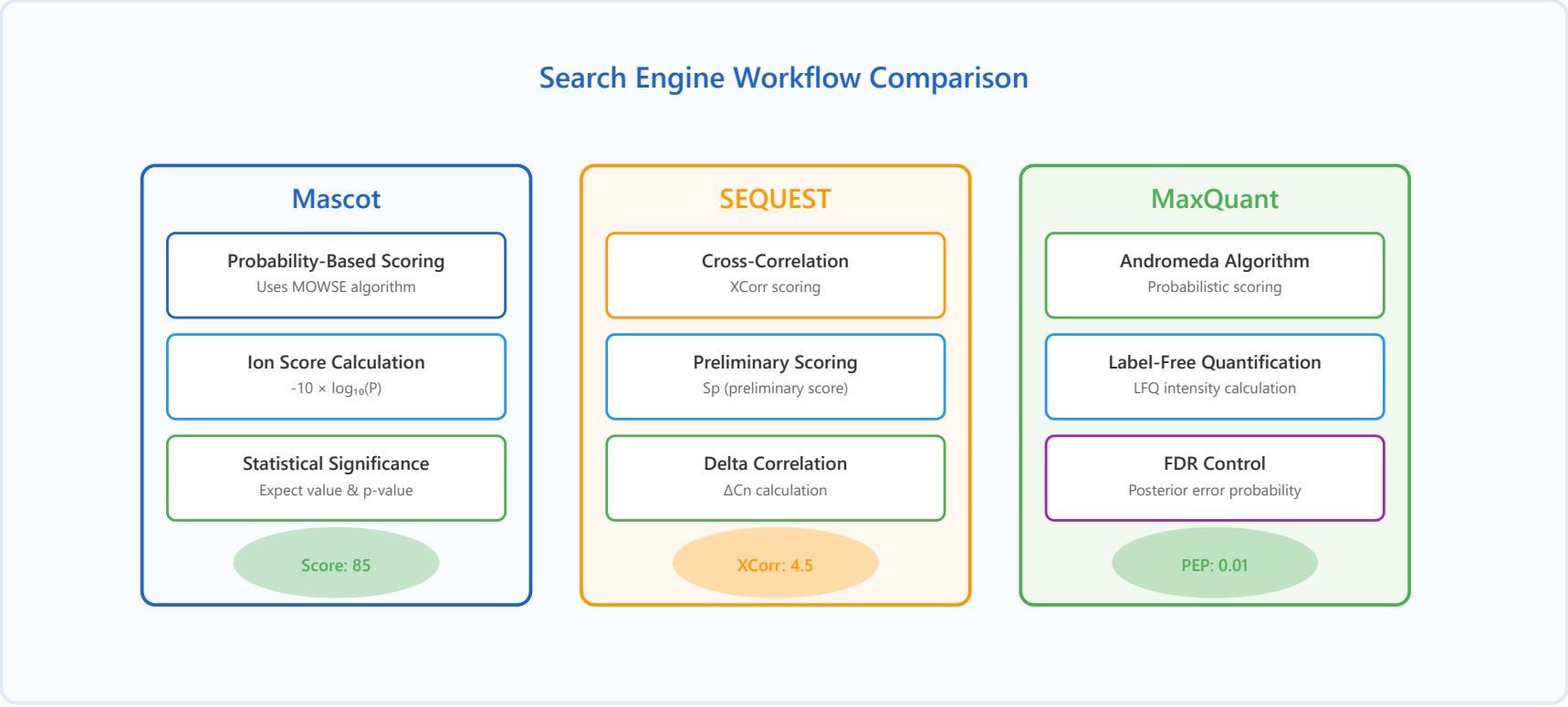
- Fixed modifications (e.g., carbamidomethylation)
- Variable modifications (e.g., oxidation)
- Balance between sensitivity and specificity

1. Search Engines

Computational tools for matching experimental spectra to theoretical sequences

Overview

Proteomics search engines are specialized software tools that match experimental MS/MS spectra against theoretical spectra derived from protein databases. Each search engine employs unique algorithms and scoring systems to identify peptide sequences from mass spectrometry data.



Key Search Engines

Search Engine	Scoring Method	Key Features	Best Use Case
Mascot	Probability-based (MOWSE)	Ion score, Expect values	Standard shotgun proteomics
SEQUEST	Cross-correlation (XCorr)	XCorr, ΔCn, Sp score	High-resolution MS data
MaxQuant	Andromeda algorithm	LFQ, SILAC quantification	Quantitative proteomics
X!Tandem	Hyperscore	Open-source, fast	Large-scale datasets

📊 **Practical Example: Score Interpretation**

For a peptide AGFAGDDAPR identified from a spectrum:

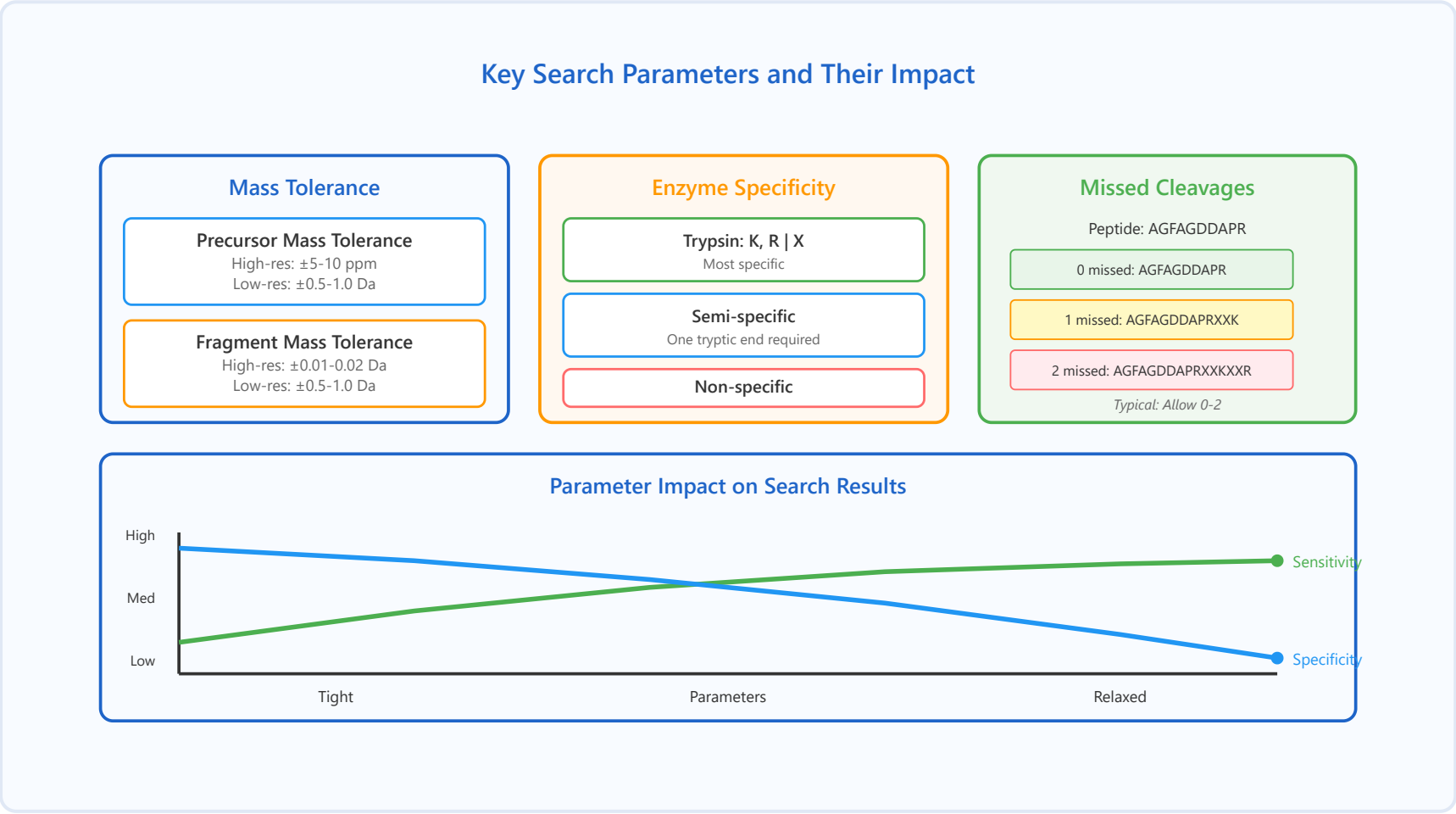
- **Mascot Score: 65** (threshold ≥ 30 for $p < 0.05$)
- **SEQUEST XCorr: 3.8** (charge +2, good match ≥ 2.5)
- **MaxQuant PEP: 0.005** (0.5% probability of false match)

2. Parameter Optimization

Fine-tuning search parameters for optimal identification accuracy

Overview

Parameter optimization is critical for maximizing the sensitivity and specificity of peptide identification. Proper parameter settings directly impact the number and quality of peptide-spectrum matches (PSMs) obtained from database searches.



Critical Parameters

- **Mass Tolerance:** Defines the acceptable mass deviation between experimental and theoretical values. High-resolution instruments (Orbitrap, Q-TOF) allow tighter tolerances (±5-10 ppm for precursor, ±0.01-0.02 Da for fragments), improving specificity.
- **Enzyme Specificity:** Specifies digestion patterns. Fully specific searches (both termini must match enzyme cleavage sites) are faster but may miss incomplete digestions. Semi-specific or non-specific searches increase sensitivity but reduce specificity.
- **Missed Cleavages:** Accounts for incomplete enzymatic digestion. Allowing 0-2 missed cleavages is standard, balancing search space expansion with computational efficiency.
- **Charge States:** Specifies the range of precursor charge states to consider (typically +2 to +4 for tryptic peptides). Incorrect charge state assignment leads to failed identifications.

✓ Best Practices

- Start with recommended instrument-specific parameters
- Calibrate mass accuracy using known peptides
- Balance sensitivity vs. search time based on dataset size
- Use diagnostic plots to verify parameter appropriateness

Example Configuration

Orbitrap Q Exactive settings:

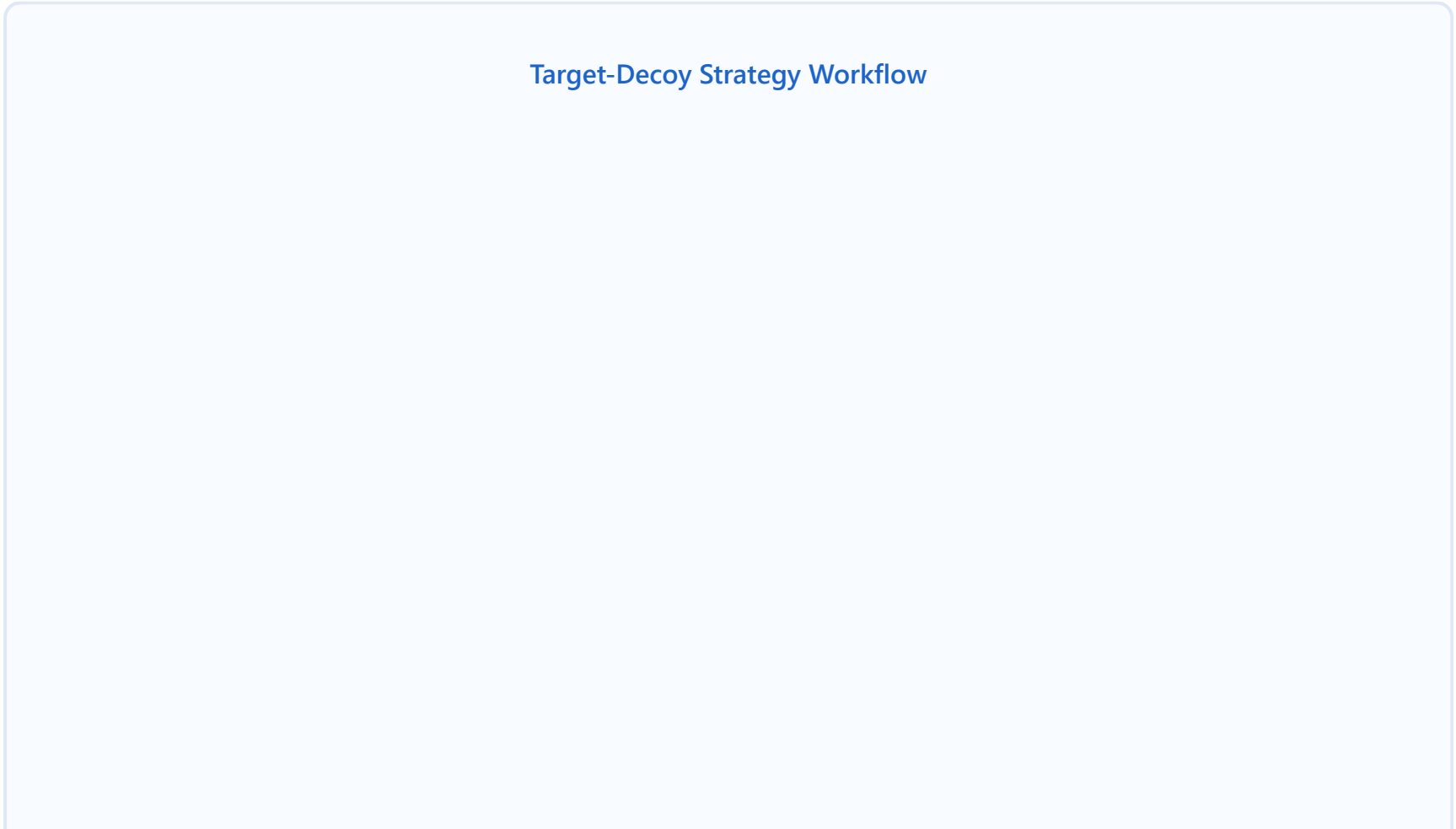
- Precursor tolerance: ±10 ppm
- Fragment tolerance: ±0.02 Da
- Enzyme: Trypsin/P (cleaves after K/R, not before P)
- Missed cleavages: 2
- Charge states: +2, +3, +4

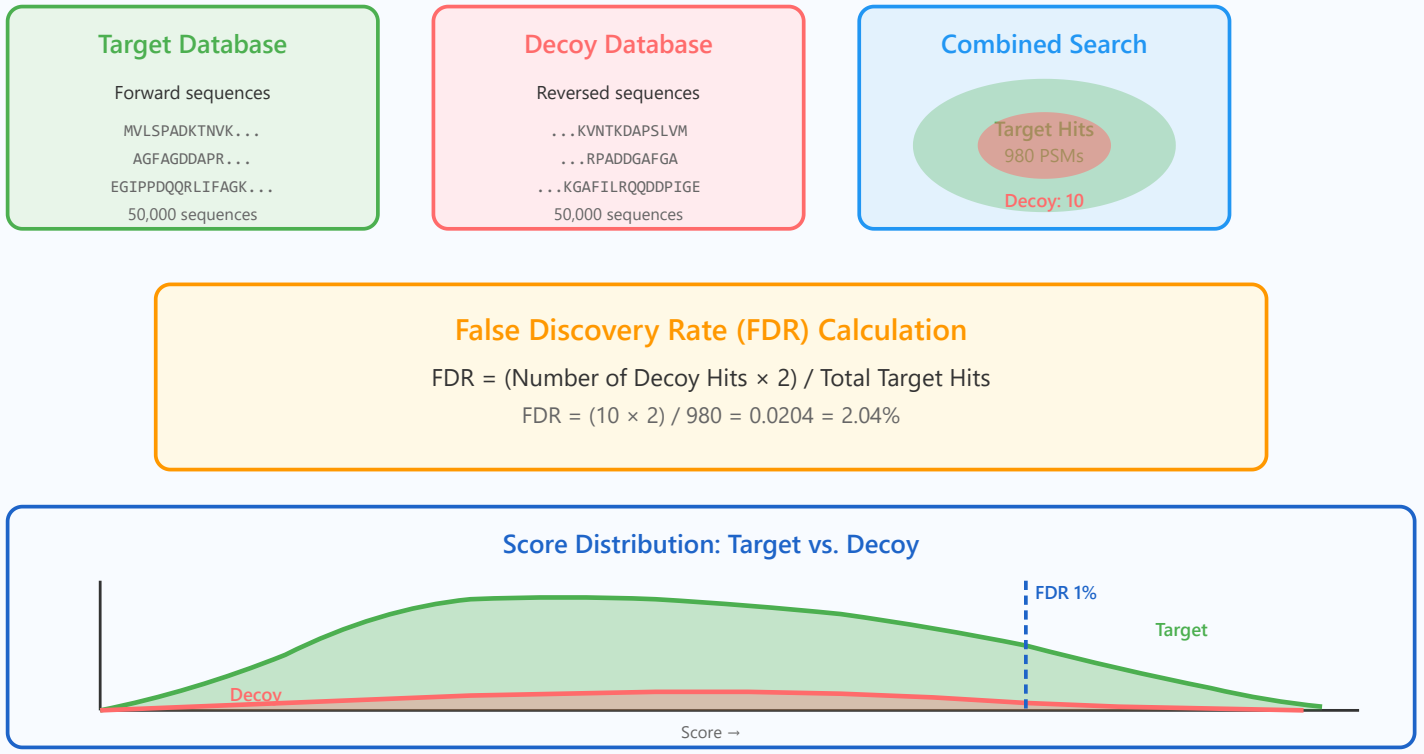
3. Decoy Databases

Statistical validation through target-decoy approach

Overview

Decoy databases are artificial protein databases used to estimate the false discovery rate (FDR) in proteomic experiments. By matching experimental spectra against both real (target) and artificial (decoy) sequences, researchers can statistically validate their identifications and control the error rate.





Decoy Generation Methods

- **Reversed Sequences:** Most common approach. Each protein sequence is reversed while maintaining the same amino acid composition. Example: AGFAGDDAPR → RPADDGAFGA. This preserves mass distribution but creates non-biological sequences.
- **Shuffled Sequences:** Randomly shuffle amino acids within each protein while maintaining terminal residues (for enzyme specificity). Provides more randomization than reversal but requires careful implementation to avoid creating target sequences.
- **Pseudo-Reversed:** Reverses sequences while keeping terminal K/R residues in place to maintain tryptic cleavage patterns, improving decoy quality for enzyme-specific searches.

📈 Interpreting FDR Results

Dataset Example:

- Target hits at score threshold: 980 PSMs
- Decoy hits at score threshold: 10 PSMs
- Estimated FDR = (10 × 2) / 980 = 2.04%

At 1% FDR threshold: Reduce cutoff to accept ~500 target PSMs with ~5 decoy hits

✓ Quality Control Metrics

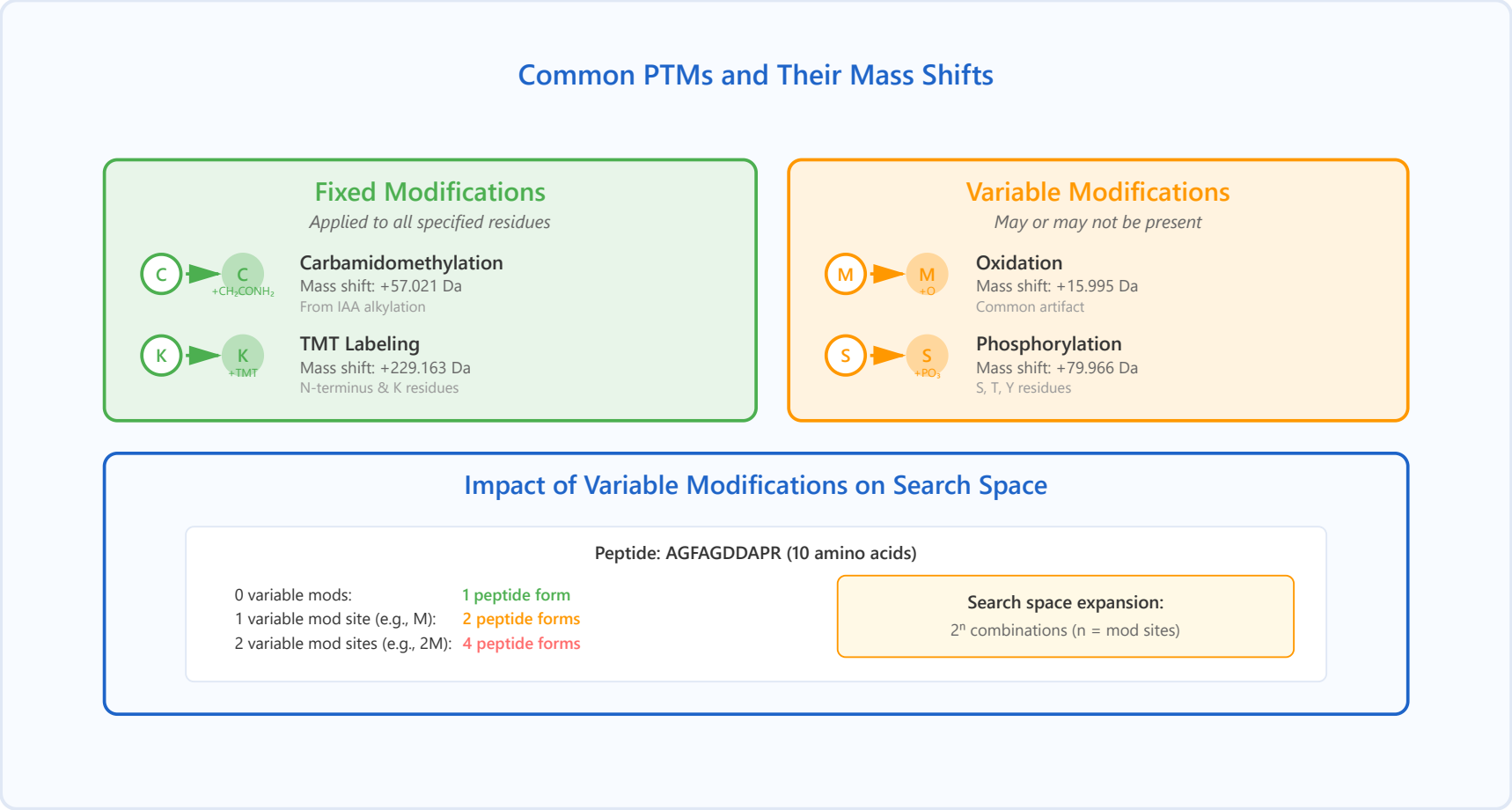
- **PSM-level FDR:** Controls false matches at spectrum level (typically 1%)
- **Peptide-level FDR:** Controls unique peptide sequences (1-5%)
- **Protein-level FDR:** Controls protein inference errors (1-5%)
- Use hierarchical FDR: Most stringent at PSM level, relaxed at protein level

4. Post-Translational Modifications (PTMs)

Identifying and characterizing chemical modifications in proteins

Overview

Post-translational modifications are chemical changes to proteins that occur after translation. These modifications regulate protein function, localization, and interactions. In proteomics database searching, PTMs are specified as mass shifts at specific amino acids, significantly expanding the search space and complexity.



Types of Modifications

Modification Type	Target Residue	Mass Shift (Da)	Biological Function
Phosphorylation	S, T, Y	+79.966	Signal transduction, regulation
Acetylation	K, N-term	+42.011	Gene regulation, protein stability
Methylation	K, R	+14.016 (mono)	Transcription regulation
Ubiquitination	K	+114.043 (Gly-Gly)	Protein degradation, signaling
Oxidation	M, W	+15.995	Artifact or oxidative stress
Deamidation	N, Q	+0.984	Protein aging, artifact

🔬 Practical Example: Phosphopeptide Analysis

Peptide: AGFS[+80]GDDAPR
Unmodified mass: 1,000.45 Da
Phosphorylated mass: 1,080.42 Da (+79.97 Da)
Interpretation: Serine at position 4 is phosphorylated
Biological relevance: Potential kinase substrate involved in signaling

⚠️ Search Strategy Considerations

- **Limit variable modifications:** Each additional variable mod exponentially increases search space (2-3 max recommended)
- **Use two-pass searches:** First search with common mods, second search with expanded mod set on unidentified spectra
- **Consider sample type:** Biological mods (phosphorylation) vs. chemical artifacts (oxidation)
- **Enrichment-aware:** Increase mod allowance for enriched samples (e.g., phospho-enrichment)