As QuantConnect's backtest Python file has limited functionality, we separate **portfolio design** from **trading** to preserve causality and make the backtest faithfully executable. Consequently, our project is broken down into two parts: an *offline* research layer, where we perform data analyses to generate portfolios, and an *online* execution layer, where the backtest is conducted.

The research layer runs offline and, using only information available up to an as-of date, computes a portfolio of candidate set and optimal long-only weights. These, along with other relevant parameters (described below), are written as immutable snapshots on fixed release dates (1^{st}, 11^{th}, and 21^{st} of each month). The execution layer runs online, loads the latest latest snapshot strictly at the close of the designated date, and places **trade-on-close** orders so prices align with the snapshot timestamp.

Unless otherwise specified, all subsequent descriptions assume a fixed timestamp t so that the analyses are done using information already available by t. We also reserve letter i (and sometimes j) to denote a generic symbol/stock.

Goal: given any time t, build a "diverse" and strong portfolio that hopefully works well in the near future.

1 Designing the Portfolio

1.1 Universe and Data

Filtering the Universe. We restrict our investable universe to restricted to large, seasoned U.S. equities so that capacity, costs, and estimation error are all well behaved, and that measured volatilizes and correlations reflect actual tradability rather than microstructure noise. Concretely, we take the current S&P 500 constituents (at time t) and apply simple investability filters: (i) a *minimum price-level* to avoid tick-size effects, (ii) a *minimum market capitalization* to rule out marginal listings, and (iii) a *minimum average daily dollar volume (ADV)* to ensure scaling, and finally, (iv) a *minimum history length* to stabilize covariances later. If L is the ADV lookback, we compute **dollar liquidity** of stock i (at time t) as

$$\text{ADV}_{i,t} = \frac{1}{L}\sum_{\ell=1}^{L} P_{i,t-\ell}\text{Vol}_{i,t-\ell}$$

and require $\text{ADV}_{i,t} \geqslant A_{\min}$, $P_{i,t} \geqslant P_{\min}$ (price), $\text{MktCap}_{i,t} \geqslant M_{\min}$, and $N_i \geqslant N_{\min}$ trading days of history. The hyperparameters are described in somewhere. This pairing eliminates thin and immature names that would otherwise dominate risk estimates or be infeasible to transact at target weights.

All prices series are split- and divident-adjusted to obtain total returns. If $P_{i,t}$ denotes the close price and $D_{i,t}$ cash distribution on day t, we work with **returns**

$$r_{i,t} = \frac{P_{i,t} + D_{i,t}}{P_{i,t-1}} - 1, \qquad R_{i,t} = \log(1 + r_{i,t}) \tag{1}$$

per standard practice.

Data Cleaning. Accounting variables are aligned to public availability: if a fundamental $X_{i,\tau}$ is filed at τ, it cannot influence any feasure used before τ. This point-in-time alignment removes look-ahead; in practice, we forward-fill fundamentals only within a conservative window and exclude the observation otherwise from cross-sectional modeling.

Sector classifications are attached to each name and serve two roles. First, they provide taxonomy for diversification constraints in sec. Second, they pivot feature normalization so signals compare like with like rather than conflating sector premia with stock-specific effects. To keep inputs robust, we attenuate extremes and remove sector-level drifts before modeling. For any raw cross-sectional feature $x_{i,t}$ inside sector g, we winsorize to intra-sector quantiles and standardize within sector by clipping values outside 1^{st} or 99^{th} quantiles, and normalize into mean zero and unit variance. Intuitively, we wish to prevent a handful of outliers or section composition drifts from dominating rankings and regressions, thereby making expected return and risk estimates more stable out-of-sample.

Calendar and timestamp conventions follow from U.S. market hours (EST) and all series are in USD. Because the execution policy is trade-on-close, all research-time features and targets are defined at the close, and the backtest consumes those snapshots strictly at the corresponding close.

1.2 Regressing on Expected-Return Prior

The Prior. The **expected-return prior** is designed to be simple, interpretable, and importantly, *conservative*. Intuitively, we seek stocks that are *trending upward, financially strong*, and exhibit *low idiosyncratic risk* (and optimally cheap) after stripping out broad sector effects so we compare like with like. This prior is *not* meant to be a precise point forecast; instead, it supplies a stable direction that remains robust after shrinkage and downstream constraints.

For each stock t at time t, we ask four questions and define four quantities correpsondingly:

- *Is the stock trending on a medium horizon*? Formally we use the $12-1$ cumulative gross return (skipping most recent month to avoid short-term reversal), defined by the momentum

$$M_{i,t} = \prod_{m=2}^{12}(1 + r_{i,t-m}) - 1.$$

- *How profitably does the firm convert its asset base into earnings?* This term is captured by a firm's operating profitability scaled by total sets, $Q_{i,t} = \mathrm{OP}_{i,t}/\mathrm{Assets}_{i,t}$.

- *How quiet is the stock idiosyncratically after removing common factors?* Formally, the low-rist tilt is the negative of the residual volatility estimated over a rolling window L:

$$\hat{\sigma}_{\epsilon,i,t} = \left(\frac{1}{L-1}\sum_{r=t-L+1}^{t}\epsilon_{i,\tau}^2\right), \qquad L_{i,t} = -\hat{\sigma}_{\epsilon,i,t}.$$

  We note that $\epsilon_{i,\tau}$ is the idiosyncratic (residual) return defined in later and mention them here as a forward pointer.

- *How cheap are its earnings relative to the price?* we let value be the earnings yield $V_{i,t} = E_{i,t}/P_{i,t}$.

Within each sector g and at time t, we winsorize (clipping values outside 1^{st} or 99^{th} percentile) and standardize each quantity into mean zero and unit variance. We slightly abuse the notation and keep them the ssame. Then, for each stock i at time t we obtain the *sector-neutralized* signal vector

$$z_{i,t} = [z_{i,t}^{(M)}, z_{i,t}^{(Q)}, z_{i,t}^{(L)}, z_{i,t}^{(V)}].$$

The Ridge Regression. We map signals to one-period-ahead returns using a **ridge regression** trained on a rolling window, with the investable set U_τ at each historical time τ. If $r_{i,\tau+1}$ denotes the next-period total return, the fitted coefficients satisfy

$$\hat{\beta}_t = \arg\min_{\beta} \sum_{\tau=t-L+1}^{t}\sum_{i\in U_\tau}(r_{i,\tau+1} - z_{i,\tau}^\top\beta)^2 + \lambda_\beta\|\beta\|_2^2, \qquad \hat{\mu}_{i,t} = z_{i,t}^\top\hat{\beta}_t. \tag{2}$$

Because cross-sectional alpha is noisy, we shrink these fitted-returns toward zero using a James-Stein style factor (cite) that adapts to the dispersion of $\hat{\mu}$ and the realized **information content** (IC) of the signals,

$$\tilde{\mu}_{i,t} = (1-\kappa_t)\hat{\mu}_{i,t}, \qquad \kappa_t = \frac{\mathrm{Var}(\hat{\mu}_{\cdot,t})}{\mathrm{Var}(\hat{\mu}_{\cdot,t}) + \sigma_{\mathrm{IC},t}^2}. \tag{3}$$

Intuitively, when the cross-section looks "wide" but the signals have not been delivering commensurate forward rank IC, the shrinkage factor rises and the prior is pulled closer to zero; when signals are genuinely informative, shrinkage relaxes. This results in a stable, sector-neutral ranking that prefers stocks with the "good" properties we seek, while guarding against over-confident forecasts that would otherwise hurt downstream estimation.

1.3 Dual-Regime Risk Model

Stock market can be bullish or bearish (or sideways). In our analysis, we partition them into two regimes: *calm* (good) or *stress*ful (bad). We need to build a **risk model** that is credible in both conditions so that diversification measured at research time survives when correlations arise. Intuitively, we separate broad, shared risk from stock-specific noise, and we then estimate the two covariances, one for calm markets and one for stressful markets, before shrinking both toward structured, well-conditioned candidates. This avoids overfitting the sample covariance and prevents "fake diversification" that disappears exactly when it is needed the most.

Decomposition the Returns. Formally, we decompose each stock's return (Equation (1)) into **factor** and **idiosyncratic** components, $r = Bf + \epsilon$ (in vector forms) where r is the return, f the vector of factor returns (e.g. market, sectors), B the matrix of factor exposures (the **beta**s), and ϵ the residual (idiosyncratic) returns. We estimate B by a ridge regression on the last L trading days to stabilize loadings,

$$\hat{B} = \arg\min_{B} \sum_{\tau=t-L+1}^{t} \|r_\tau - Bf_\tau\|_2^2 + \lambda_B \|B\|_F^2, \qquad \hat{\epsilon}_\tau = r_\tau - \hat{B} f_\tau.$$

Intuitively, Bf captures the broad movements we cannot diversity away (e.g. market shocks), while ϵ is the stock-level noise we *can* (and will) diversify.

The Two Regimes. We define the two regimes based on a market-state indicator built from SPY. Let ν_τ be SPY's 21-day rolling volatility by time τ. Over the same trailing window $[t - L + 1, t]$, we compute the lower and upper terciles (33rd and 67th percentiles) of $\{\nu_\tau\}$. Dates with $\nu_\tau \leqslant$ the lower tercile form the **calm dates** C; dates with ν_τ above the *upper* tercile form the **stress dates** $\mathcal{S}$. Note that they do not partition all dates; we leave the middle tercile unused on purpose for regime-specific estimation. We next estimate the covariance of returns within C or $\mathcal{S}$ using $r = Bf + \epsilon$; written compactly in linear algebraic terms, we let

$$\Sigma^{(C)} = \hat{B}\Sigma_f^{(C)}\hat{B}^\top + D^{(C)} \qquad \Sigma^{(\mathcal{S})} = \hat{B}\Sigma_f^{(\mathcal{S})}\hat{B}^\top + D^{(\mathcal{S})} \tag{4}$$

where $\Sigma_f^{(k)} = \mathrm{Cov}(f_\tau)_{\tau\in k}$ and $D^{(k)} = \mathrm{diag}(\mathrm{Var}(\hat{\epsilon}_\tau)_{\tau\in k})$. We further apply Ledoit-Wolf (cite) on the Σ's to make them well-conditioned; we write the final results as $\Sigma^{(\text{calm})}$ and $\Sigma^{(\text{stress})}$. In parallel, we also retain the residual correlation matrices, computed from $\hat{\epsilon}_\tau$ within each regime; we denote them by $R^{(\text{calm})}$ and $R^{(\text{stress})}$.

In short, $r = Bf + \epsilon$ separates common from idiosyncratic movement; the two calm vs. stress regimes give two views of risk; applying shrinkage keeps both views well-conditioned; and we carry forward the residual correlations for diversification and total covariances for allocating weights in the next section.

1.4 Candidate Selection

We now seek a set K of good candidate stocks that are individually attractive and non-redundant in *both* regimes. Concretely, we seek stocks whose expected return per unit of idiosyncratic risk is high, and we avoid picking several stocks that tend to move together once the broad market effects are removed.

At time t, we first form a calm-regime *quality score* by dividing the shrunk prior (Equation (3)) by the calm idiosyncratic volatility, $q_i = \tilde{\mu}_i \cdot (\mathrm{Var}(\epsilon_i)_{\text{calm}})^{-1/2}$. We pre-select top m names by $\{q_i\}$ and form a pool. We then choose a set S of size K (a hyperparameter) by maximizing a quality-diversity objective

$$\max_{S\subset\text{pool}|S|=K} \lambda \sum_{i\in S} q_i + (1-\lambda)\left[\beta \log\det(R^{(\text{calm})}[S]) + (1-\beta)\log\det(R^{(\text{stress})}[S])\right] \tag{5}$$

subject to sector cap constraints (hyperparemeters). The first term $\sum_i q_i$ favors stocks with strong idiosyncratic risk-adjusted priors, whereas the two log-determinant terms favor sets whose residuals span many independent directions in both regimes. In practice (and in our implementation), we solve this combinatorial problem greedily: starting empty and iteratively add any feasible additional stock that yields the largest marginal increase until we reach K total.

1.5 Portfolio Construction

Let S be the resulting set from the previous section. We now need to weight the portfolio appropriately by balancing expected return against *worst-case* risk across the two risk regimes, while enforcing realistic trading and diversification constraints. Recall $\Sigma^{(\text{calm})}$ and $\Sigma^{(\text{stress})}$ from red; we further strict them to S without changing notation. For each scenario $k \in \{\text{calm}, \text{stress}\}$, define a corresponding volatility $\sigma_k(w) = (w^\top \Sigma^{(k)} w)^{1/2}$. Let w_{t-1} be the last period's weight on S. We wish to maximize $\tilde{\mu}^\top w$, the expected return. To abide by the additional requirements, we impose the additional regularization:

- a penalty bounded from below by the volatilizes from both regimes;

- a penalty against spiky allocations (that put too much weight on one stock);
- a liquidity-aware trading cost that penalizes moving weight into less liquid names.

These can be translated into a nonlinear program, defined as follows. In the objective, the three penalty terms are listed in the order they are introduced. B is the matrix of stock exposures from red (restricted to S), and Γ is the liquidity-scaled matrix with $\Gamma_{i,i} \propto 1/\mathrm{ADV}_i$. Hyperparameters are $\gamma, \eta_w, w_{\max}, [\ell, u_g]$ for each sector g, c_B (exposure), $H_{\max}$ (HHI cap), and $\mathrm{TO}_{\max}$ (turnover). We currently assume that the portfolio is fully invested and will address additional risk management in the actual backtest. We now present the maximization program over $w \in \mathbb{R}^K$.

$$
\begin{aligned}
\text{maximize} \quad & \tilde{\mu}^\top w - t - \eta_w \|w\|_2 - \frac{1}{2}(w - w_{t-1})^\top \Gamma (w - w_{t-1}) \\
\text{subject to} \quad & t \geqslant \gamma \cdot \sigma_{\text{calm}}(w), \qquad t \geqslant \gamma \cdot \sigma_{\text{stress}}(w) && \text{(worst-case risk)} \\
& \mathbf{1}^\top w = 1, \qquad 0 \leqslant w_i \leqslant w_{\max} \text{ for each } i \in S && \text{(weight constraints)} \\
& \ell_g \leqslant \sum_{i \in S \cap g} w_i \leqslant u_g \text{ for each sector } g && \text{(per-sector requirements)} \\
& \|B^\top w\|_\infty \leqslant c_F, \qquad \sum_{i \in S} w_i^2 \leqslant H_{\max} && \text{(exposure \& concentration)} \\
& \|w - w_{t-1}\|_1 \leqslant \mathrm{TO}_{\max} && \text{(turnover budget)}
\end{aligned}
$$

We solved this convex second-order program on the K-stock subspace with a numerical solver. This completes the section of portfolio design. Before backtesting, we will iteratively repeat this process approximately every 10 days (on the 1st, 11th, and 21st of each month) in the backtest window and store the results. Then, the backtest engine will retrieve these configurations at appropriate time without producing lookahead leakage.

2 Executing the Backtest