PARSE

Peer-Augmented Residual Signal Ensemble

Jonas Bock Codex

University of Münster

October 2026

Abstract

We construct equity portfolios by forecasting standardized residual returns across three risk books. Each book combines a LightGBM forecast on 402 firm characteristics with a second forecast augmented by 402 peer-relative characteristics. The books hedge return factors together with industry, characteristic-PCA or affine exposures. A fixed 50/50 mixture combines the raw and peer-augmented branches after risk normalization. Over January 1990 to December 2023, the resulting portfolio achieves an annualized gross Sharpe ratio of 4.7371.

1 Method

The model builds on the residual-book framework of Gehling (2026) and the peer-relative characteristic construction of Qi et al. (2026). The forecasting problem is to distinguish stocks that outperform comparable firms after common return components are removed. Peer features supply information about relative characteristics, while an affine representation supplies a separate set of common exposures. We use equal weights within each three-book branch and a fixed weight of 50% on the peer-augmented branch.

1.1 Peer-relative characteristics

Let $X_t \in \mathbb{R}^{n_t \times 402}$ contain the supplied characteristics at month-end t. The raw branch uses X_t directly. For the augmented branch, let $q_{ij,t}$ be the within-month percentile rank of characteristic j, centered by $1/2$. Standardized ranks are projected onto up to 20 principal components. Ward clustering forms

$$K_t = \min\{n_t, \max[10, \min(60, \operatorname{round}(n_t/100))]\} \tag{1}$$

peer groups. If $g_t(i)$ denotes the group containing stock i, the additional characteristic is

$$\widetilde{q}_{ij,t} = q_{ij,t} - \frac{\sum_{k \in g_t(i)} \mathbf{1}_{\{q_{kj,t} \text{ observed}\}} q_{kj,t}}{\sum_{k \in g_t(i)} \mathbf{1}_{\{q_{kj,t} \text{ observed}\}}}. \tag{2}$$

Missing ranks are filled by zero for clustering and omitted from peer means; missing augmented features remain missing for LightGBM. The 804-input branch concatenates X_t and $\widetilde{q}_t$. Groups are estimated from the entire current-month characteristic cross-section before trading-history eligibility is applied. No returns enter this step.

1.2 Risk estimates and residual books

Each month uses up to $T = 252$ preceding trading days. A stock must have at least 200 observed daily excess returns and positive variance; at least 100 eligible stocks are required. Daily returns are demeaned and normalized stock by stock, with missing observations filled by zero after demeaning. Principal components above the Marchenko–Pastur edge $(1 + \sqrt{n_t/T})^2$ form return loadings B_t. Define

$$\sigma_{i,t} = \sqrt{21}\ \operatorname{sd}(r_{i,d} : d \le t), \qquad \psi_{i,t} = \max\left\{1 - \sum_k B_{ik,t}^2,\ 0.05\right\}, \tag{3}$$

where the volatility uses the same trailing sample and its observed daily returns. Write $\Sigma_t = \mathrm{diag}(\sigma_t)$ and $\Psi_t = \mathrm{diag}(\psi_t)$. The three exposure matrices are

$$G_{1,t} = [B_t,\, \Sigma_t^{-1}\mathbf{1},\, \Sigma_t^{-1}D_t], \tag{4}$$

$$G_{2,t} = [B_t,\, \Sigma_t^{-1}\mathbf{1},\, C_t], \tag{5}$$

$$G_{3,t} = [B_t,\, \Sigma_t^{-1}\mathbf{1},\, \beta(Z_t)]. \tag{6}$$

Here D_t contains two-digit SIC industry indicators. Industries with fewer than ten stocks and stocks without a code form one group. The columns of C_t are principal components of current ranked characteristics above their own Marchenko–Pastur edge. The third matrix uses the affine loadings defined below. All books are return-factor and dollar neutral; their additional restrictions apply to their respective books.

1.3 Affine representation

For each historical formation month, Z_t contains centered characteristic ranks, with missing ranks filled at the median, and $y_{t+1} = \Sigma_t^{-1} r_{t+1}$. Form the ridge coefficient vector

$$h_t = (Z_t'Z_t + 0.001 n_t I)^{-1} Z_t' y_{t+1}. \tag{7}$$

An affine loading map and a linear factor map reconstruct returns as

$$\beta(Z_t) = Z_t A + \mathbf{1}a', \qquad f_t = F h_t, \qquad \widehat{y}_{t+1} = \beta(Z_t) f_t, \tag{8}$$

with five factors. The representation minimizes mean squared reconstruction error using Adam, learning rate 0.001, weight decay 0.0001 and seed 42. It is estimated from 384 historical months with formation dates through November 1983. All associated returns are therefore available by December 1983. Epoch selection uses formation dates before December 1978 for training and the remaining historical dates for validation, with at most 80 epochs, patience 10 and minimum improvement 10^{-7}. Eight epochs are selected on the supplied data, followed by a fresh fit on all 384 months. The loading map is then frozen. Before its fixed historical availability date, the implementation uses the two base books.

2 Forecasts and portfolio construction

For book b, let $U_{b,t}$ be an orthonormal basis of $\Psi_t^{-1/2} G_{b,t}$ and define the residual projection $P_{b,t} = I - U_{b,t} U_{b,t}'$. Its standardized residual target is

$$z_{b,t+1} = P_{b,t} \Psi_t^{-1/2} \Sigma_t^{-1} r_{t+1}. \tag{9}$$

For regression, each month's targets are divided by their cross-sectional standard deviation and clipped at ± 4. Each book has two LightGBM regressors (Ke et al., 2017): one on the 402 raw characteristics and one on the 804 augmented inputs.

All learners are refit annually in December using the complete expanding history of labeled formation months strictly before the refit date. Their next-month returns have been observed by that date. The latest $\min\{60, \lfloor H/2 \rfloor\}$ historical months form validation, where H is the number of available months; at least 36 historical months are required. Learning rate is 0.05, the number of leaves is 31, and minimum leaf size is the median monthly stock count in the corresponding training sample. Boosting selects at most 2,000 rounds with early-stopping patience 200 and seed 42. The selected number of rounds is refit from scratch on the entire available history, including validation.

Early stopping maximizes the historical monthly residual-book Sharpe criterion. Industry and affine books use projected-score portfolio returns. The characteristic-PCA book retains the inherited proxy $s'z / \|s\|_2$. This training criterion is distinct from the realized test-period returns reported below.

For forecast scores $s_{a,b,t}$ from branch $a \in \{\mathrm{raw}, \mathrm{peer}\}$ and book b, set

$$p_{a,b,t} = P_{b,t} s_{a,b,t}, \qquad w_{a,b,t} = \Sigma_t^{-1} \Psi_t^{-1/2} \frac{p_{a,b,t}}{\|p_{a,b,t}\|_2}. \tag{10}$$

Define the specific-risk normalization

$$\mathcal{N}_t(w) = \frac{w}{\left\| \Sigma_t \Psi_t^{1/2} w \right\|_2}. \tag{11}$$

The branch portfolios and final portfolio are

$$u_{a,t} = \mathcal{N}_t\left(\frac{1}{3}\sum_{b=1}^{3} w_{a,b,t}\right), \tag{12}$$

$$w_t = \frac{0.1}{\sqrt{12}}\,\mathcal{N}_t(0.5u_{\mathrm{raw},t} + 0.5u_{\mathrm{peer},t}). \tag{13}$$

The factor $0.1/\sqrt{12}$ scales specific risk; realized total volatility need not equal 10%. Ineligible scored stocks receive zero weights. Portfolios are rebalanced monthly and allow leverage and short positions.

3 Results

Table 1 reports realized gross performance over the 408 test months and subperiods of the Common Task Framework (Jensen et al., 2026). Annualized Sharpe is $\sqrt{12}\,\overline{R}/\,\mathrm{sd}(R)$ using sample standard deviation. The supplied portfolio weights retain the causal normalization in equation (13); their full-period annualized volatility is 24.4904%. For presentation, each table row is independently rescaled by a constant to 10% realized volatility. Sharpe is invariant to this positive constant.

Table 1: Gross performance at 10% annualized volatility.

Return period	Months	Sharpe	Annualized mean return	Max. DD
1990–2023	408	4.7371	47.37%	-5.76%
1990–1999	120	9.4037	94.04%	0.00%
2000–2009	120	5.5757	55.76%	-1.33%
2010–2023	168	3.5108	35.11%	-7.99%
2009–2023 (15 years)	180	3.5484	35.48%	-7.54%
2014–2023 (10 years)	120	2.8973	28.97%	-7.68%
2019–2023 (5 years)	60	2.2017	22.02%	-7.33%

Annualized mean return is 12 times the average monthly excess return. It uses the same annualization convention as the Sharpe ratio; at 10% volatility it equals 0.10 times Sharpe. Maximum drawdown is measured from an excess-return index initialized at one. Each row uses its own constant reporting scale; all figures are gross of trading, borrowing, financing and market-impact costs.

Table 2: Branch comparison, January 1990–December 2023.

Portfolio	Annualized gross Sharpe
Raw branch: equal thirds across three risk books	4.644319
Peer-augmented branch: equal thirds	4.658253
PARSE: fixed 50/50 combination	4.737096

The 50/50 combination improves Sharpe by 0.092776 relative to the raw equal-thirds reference. Table 1 also shows substantial variation over time.

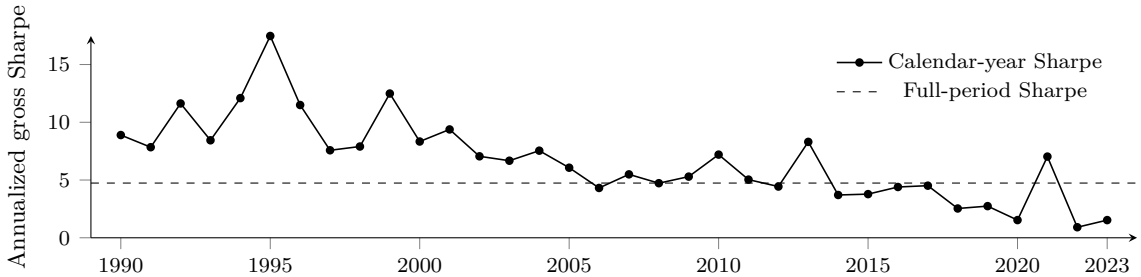


Figure 1: Calendar-year Sharpe ratios. Each point uses exactly twelve monthly excess returns; the dashed line is the full-period Sharpe. Annual estimates are descriptive and subject to substantial sampling variation.

4 Implementation and research protocol

The accompanying Python file implements `main(chars, features, daily_ret)` and returns `id`, `eom` and `w`. It covers all 885,698 scored security-months in the supplied data. The CSV is 34,827,255 bytes. Exact dependency versions are provided in `requirements.txt`. All learned states are computed from the input data frames; the model contains no embedded fitted state or external data retrieval. CPU execution uses three simultaneous four-thread learners. The random seeds, full expanding history and complete selection budgets are preserved.

References

Gehling, F. (2026). *Residual Books with Industry Neutrality* and the earlier Residual Books (LightGBM) implementation. Global Factor Data Common Task Framework. Model and documentation.

Qi, A., Kwauk, M., Yang, B., and Chen, H. (2026). *Peer-Relative Boosted Forecasts in a 414-Factor Risk Model.* Global Factor Data Common Task Framework. Methodology.

Ke, G., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30. Paper.

Jensen, T., Kelly, B., and Pedersen, L. (2026). *Global Factor Data Common Task Framework: Rules and Submission Specification.* Revision 29 September 2026, accessed 23 October 2026. Rules.