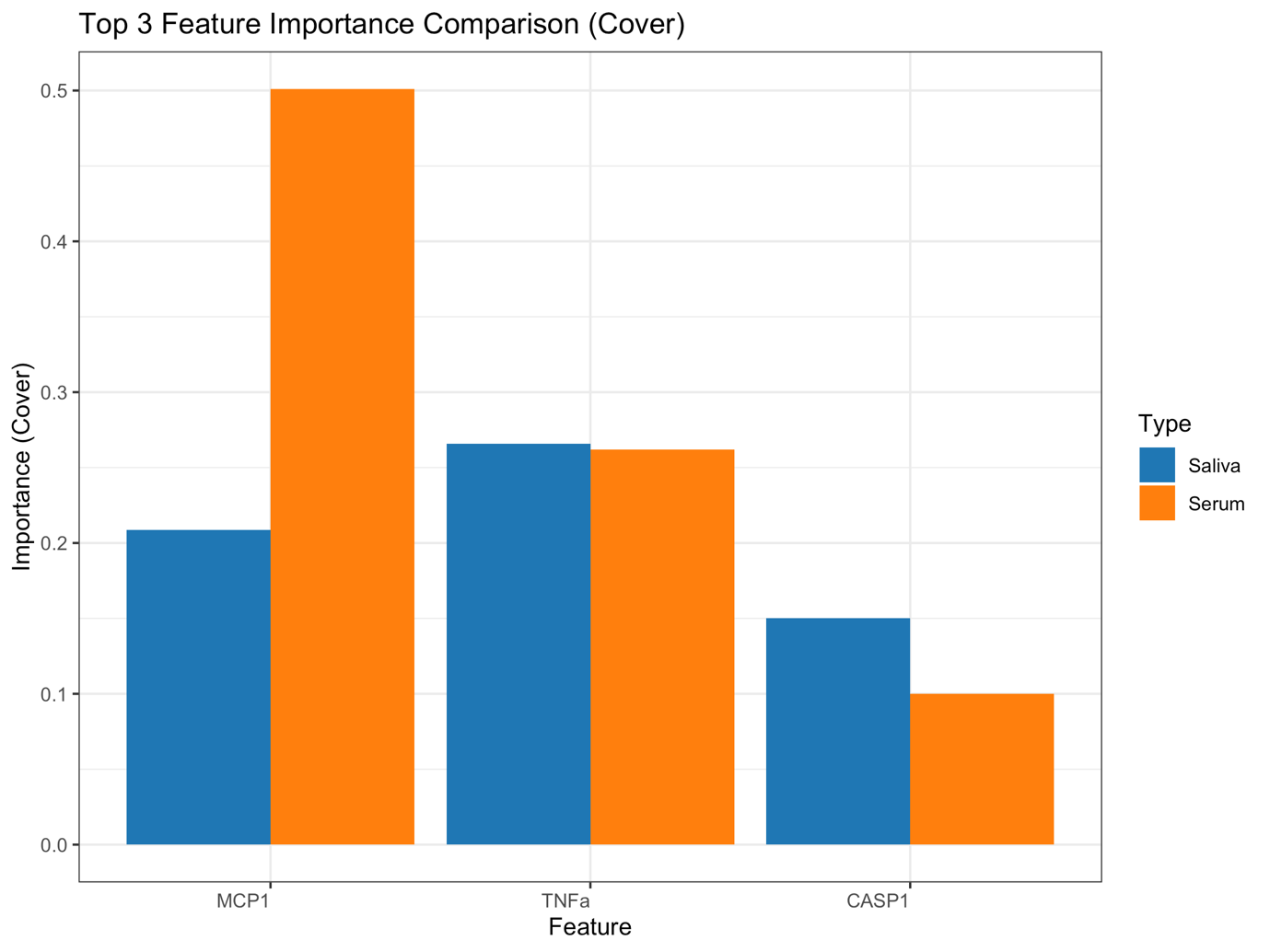
### 因子筛选与建模总结

#### 1. 初始因子选择

我们最初有 21 个唾液因子和 21 个血清因子，分别使用 XGBoost 训练模型，模型的准确率都为 0.75，并基于特征重要性矩阵筛选出了 5 个唾液因子和 3 个血清因子。对重叠的三个因子进行比较：

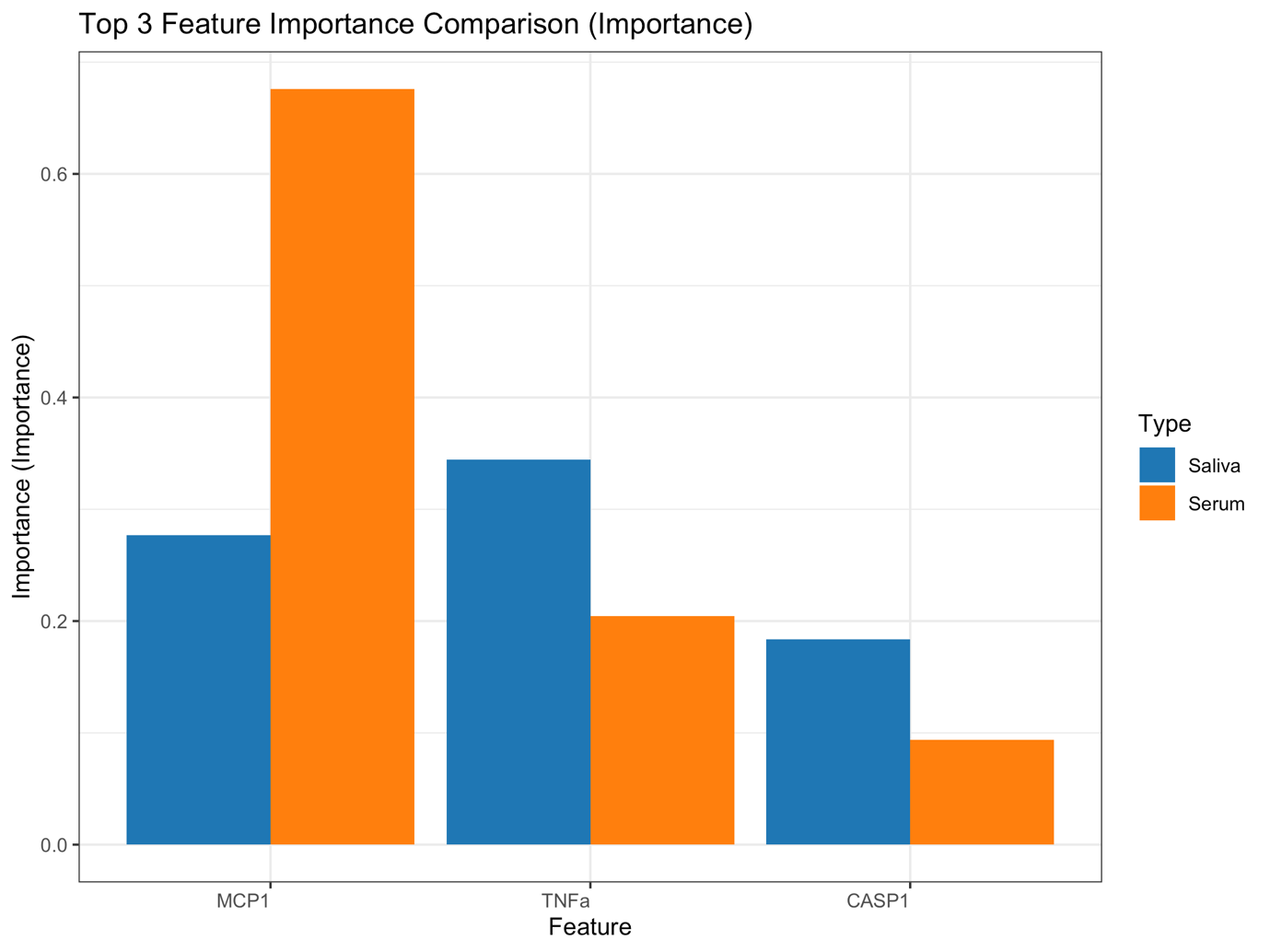


A graph with blue and orange bars

AI-generated content may be incorrect.

A graph with blue and orange bars

AI-generated content may be incorrect.



#### 2. XGBoost 模型评估

对于构建的 XGBoost 模型，并进行模型验证，F1-score 均为 0.7778，表现良好。随后，我们进行交叉验证，互换数据集进行预测，以评估数据来源的影响：

* 使用唾液因子训练 XGBoost 模型，并用血清数据测试，计算 F1score等评估指标。
* 使用血清因子训练 XGBoost 模型，并用唾液数据测试，计算 F1score等评估指标。

交叉测试结果分析：

整体来看，血清因子模型表现较优，而唾液因子的分类能力一般（Accuracy = 0.5）。由于血清模型可能存在过拟合，我们采用 5 折交叉验证重新评估 AUC 和 F1-score。

#### 3. 交叉验证评估

* 评价指标

RMSE（均方根误差）：衡量模型预测值与实际值的差异，值越小越好。

R-squared（R² 值）：反映模型拟合优度，值越接近 1 越好。

MAE（平均绝对误差）：衡量预测值与实际值的平均绝对差异，值越小越好。

* 模型表现
  + 在多种超参数组合下，RMSE 值较小，R² 值接近 1，模型拟合较好。
  + 例如，在 eta = 0.3, max\_depth = 1, colsample\_bytree = 0.6, subsample = 0.875 时，RMSE 和 MAE 均较小，表明模型性能较优。
  + RMSE 和 R² 差距较小，未出现明显过拟合。

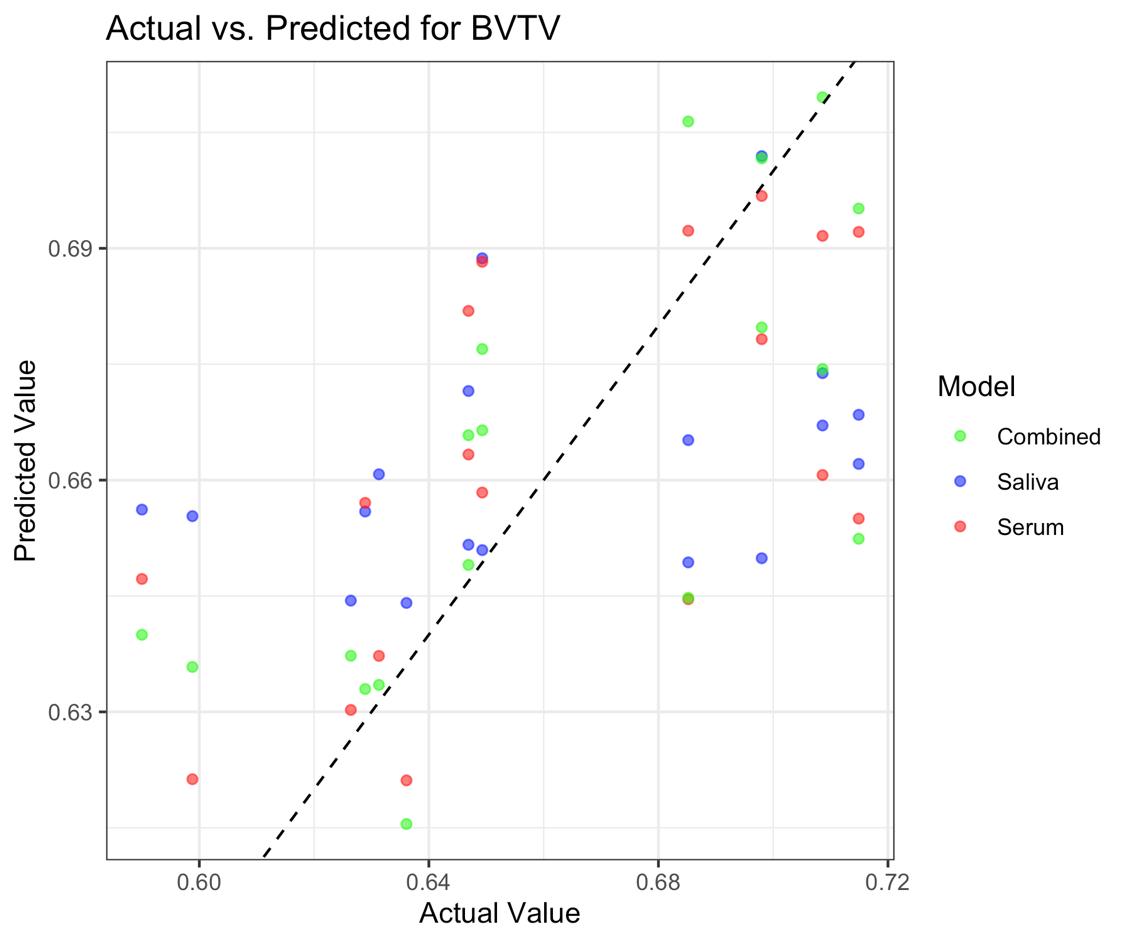
#### 4. 线性回归验证

在此基础上，我们提取 8 个因子进行线性回归分析，并对不同组合的预测能力进行比较。

* 单独使用血清因子（xueqing）相比唾液因子（tuoye）表现更优（MSE 更低，R² 更高），尤其在 BV/TV 和 Tb.Sp 变量上：  
  + BV/TV: 血清因子 R² = 0.39，高于唾液因子的 0.14。
  + Tb.Sp: 血清因子 R² = 0.95，高于唾液因子的 0.82。
  + M1: 唾液因子 R² = 0.75，高于血清因子的 0.69，可能表明 M1 变量与唾液因子的关系更强。
* 合并因子（Combined）后，整体预测效果得到提升：  
  + 所有目标变量的 MSE 降低，R² 提高。
  + BV/TV: R² 提升至 0.50。
  + Tb.Sp: R² 提升至 0.96。

### 结论

1. 血清因子在 BV/TV 和 Tb.Sp 的预测能力更强。
2. 唾液因子在 M1 的预测能力更强，可能与其生物学特性有关。
3. 合并因子后预测效果最佳，建议使用联合模型以提升预测能力。

A graph with colored dots and lines

AI-generated content may be incorrect.A graph with colored dots and lines

AI-generated content may be incorrect.