# High throughput tight binding calculation of electronic HOMO-LUMO gaps and its prediction for natural compounds

Sascha Thinius[a]

*Correspondence:* sascha.thinius@ifam.fraunhofer.de; sascha.thinius.87@gmail.com

**Hyperparameter optimization - optimized set of parameters for randomized search:**

GBR: {'alpha': 0.9, 'ccp_alpha': 0.0, 'criterion': 'friedman_mse', 'init': None, 'learning_rate': 0.0321079969138713, 'loss': 'squared_error', 'max_depth': 12, 'max_features': None, 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 1171, 'n_iter_no_change': None, 'random_state': None, 'subsample': 0.6917115704474215, 'tol': 0.0001, 'validation_fraction': 0.1, 'verbose': 0, 'warm_start': False}

MLPR: {'activation': 'relu', 'alpha': 0.00013539468639451956, 'batch_size': 'auto', 'beta_1': 0.9055614316573964, 'beta_2': 0.999, 'early_stopping': False, 'epsilon': 1e-08, 'hidden_layer_sizes': (300, 200, 100, 50), 'learning_rate': 'constant', 'learning_rate_init': 0.001, 'max_fun': 15000, 'max_iter': 500, 'momentum': 0.9, 'n_iter_no_change': 5, 'nesterovs_momentum': True, 'power_t': 0.5, 'random_state': None, 'shuffle': True, 'solver': 'adam', 'tol': 0.0005, 'validation_fraction': 0.1, 'verbose': False, 'warm_start': False}

**GBR Model Feature Importance:**

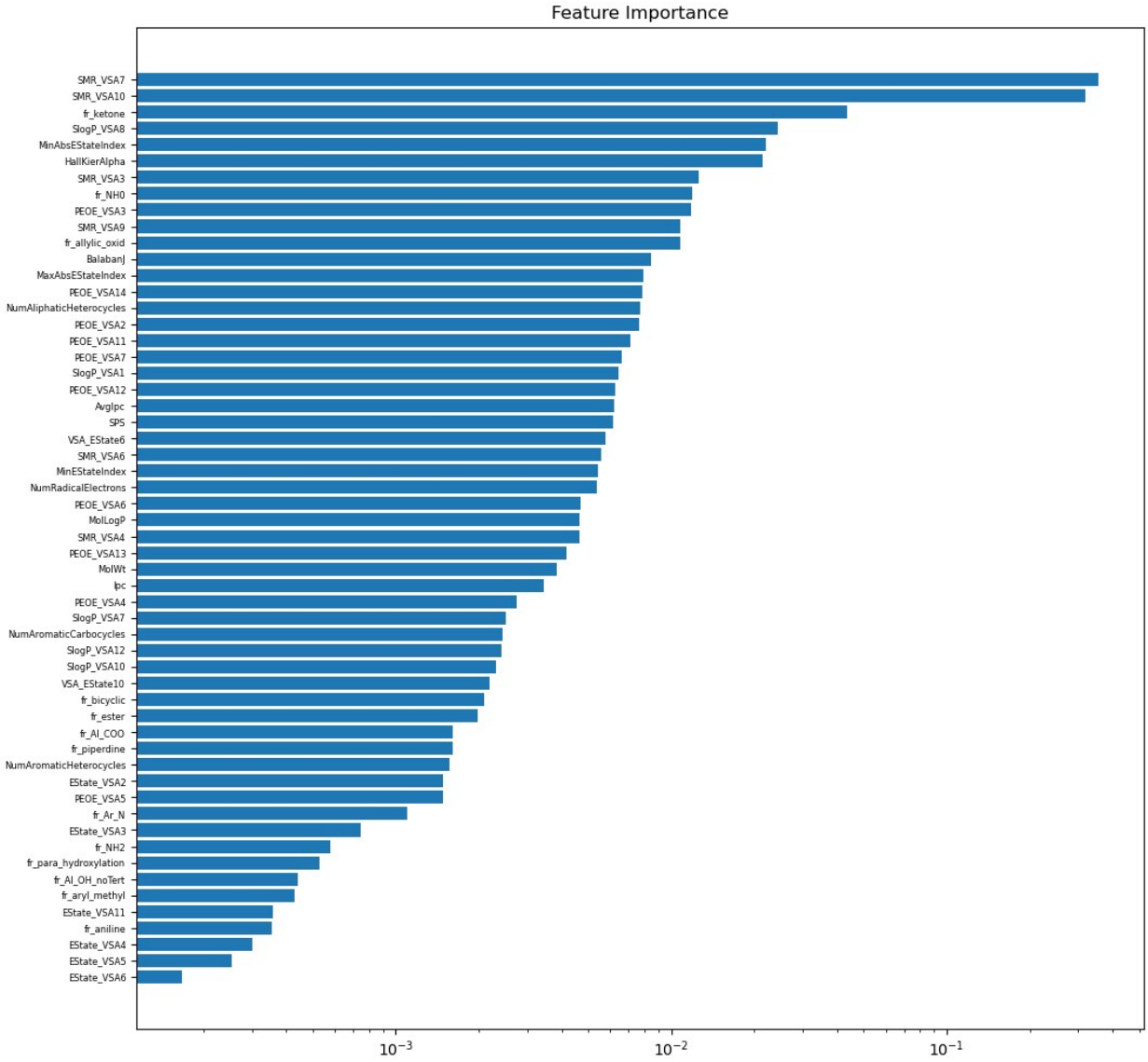


**Figure S1:** The relative importance of each molecular descriptor for the Gradient Boosting Regressor (GBR) model, sorted in descending order. The importance metric indicates the contribution of each feature to the model's predictions. The logarithmic scale on the x-axis highlights that the model's predictions are dominated by a few key features, primarily those related to molecular polarizability (SMR_VSA7, SMR_VSA10) and the presence of ketone fragments (fr_ketone).


*[a] Fraunhofer Institute for Manufacturing Technology and Advanced Materials IFAM*
*Wiener Strasse 12*
*28359 Bremen, Germany*


*Supporting Information for this article is available on the WWW under PLACEHOLDER*

**table of the single descriptor contribution for the GBR model:**

**Table S1:** The complete list of all molecular descriptors and their corresponding feature importance values for the Gradient Boosting Regressor (GBR) model, sorted in descending order.

| descriptor | feature importance |
|---|---|
| SMR_VSA7 | 0.3545370506427196 |
| SMR_VSA10 | 0.3189477603409475 |
| fr_ketone | 0.043355946009772654 |
| SlogP_VSA8 | 0.024361114189223737 |
| MinAbsEStateIndex | 0.022079855516308193 |
| HallKierAlpha | 0.02142998649717104 |
| SMR_VSA3 | 0.012575480656202306 |
| fr_NH0 | 0.011926842442014604 |
| PEOE_VSA3 | 0.011765650842991669 |
| SMR_VSA9 | 0.010827996870413808 |
| fr_allylic_oxid | 0.010791493260704166 |
| BalabanJ | 0.008404100322501391 |
| MaxAbsEStateIndex | 0.007901535542237765 |
| PEOE_VSA14 | 0.007843345900954273 |
| NumAliphaticHeterocycles | 0.007689300322622823 |
| PEOE_VSA2 | 0.0076543614597266605 |
| PEOE_VSA11 | 0.007132292603300951 |
| PEOE_VSA7 | 0.006611842332520015 |
| SlogP_VSA1 | 0.006409881164823698 |
| PEOE_VSA12 | 0.00625829887237445 |
| AvgIpc | 0.006201809042991263 |
| SPS | 0.0061684746614582555 |
| VSA_EState6 | 0.005771194866363084 |
| SMR_VSA6 | 0.005542426338198005 |
| MinEStateIndex | 0.0054140764606567925 |
| NumRadicalElectrons | 0.00537131165165876 |
| PEOE_VSA6 | 0.004669820678690445 |
| MolLogP | 0.004635016029957915 |
| SMR_VSA4 | 0.00463256729367672 |
| PEOE_VSA13 | 0.0041621834497876 |
| MolWt | 0.0038454106803835876 |
| Ipc | 0.0034334748375035337 |
| PEOE_VSA4 | 0.0027347392046638535 |
| SlogP_VSA7 | 0.002516851429005792 |
| NumAromaticCarbocycles | 0.0024361194223689506 |
| SlogP_VSA12 | 0.0024094972964538723 |
| SlogP_VSA10 | 0.0023022699682205545 |
| VSA_EState10 | 0.002183400973898018 |
| fr_bicyclic | 0.0020944360754341315 |
| fr_ester | 0.0019736029426533157 |
| fr_Al_COO | 0.0016084339785770738 |
| fr_piperdine | 0.0016024953385289705 |
| NumAromaticHeterocycles | 0.0015646714728290855 |
| EState_VSA2 | 0.0014830882240174821 |
| PEOE_VSA5 | 0.001480804233055884 |
| fr_Ar_N | 0.001104413402410501 |
| EState_VSA3 | 0.0007431130195376715 |
| fr_NH2 | 0.0005780459353178737 |
| fr_para_hydroxylation | 0.0005277236836350452 |
| fr_Al_OH_noTert | 0.0004398099196385116 |
| fr_aryl_methyl | 0.0004306268701694323 |

| | |
|---|---|
| EState_VSA11 | 0.0003571689919175693 |
| fr_aniline | 0.00035546111272848237 |
| EState_VSA4 | 0.00030027576683221797 |
| EState_VSA5 | 0.0002533586761247892 |
| EState_VSA6 | 0.00016769028112372864 |

**2D histogram plots for molecular subsets of chemical motifs (see table 3 in publication):**

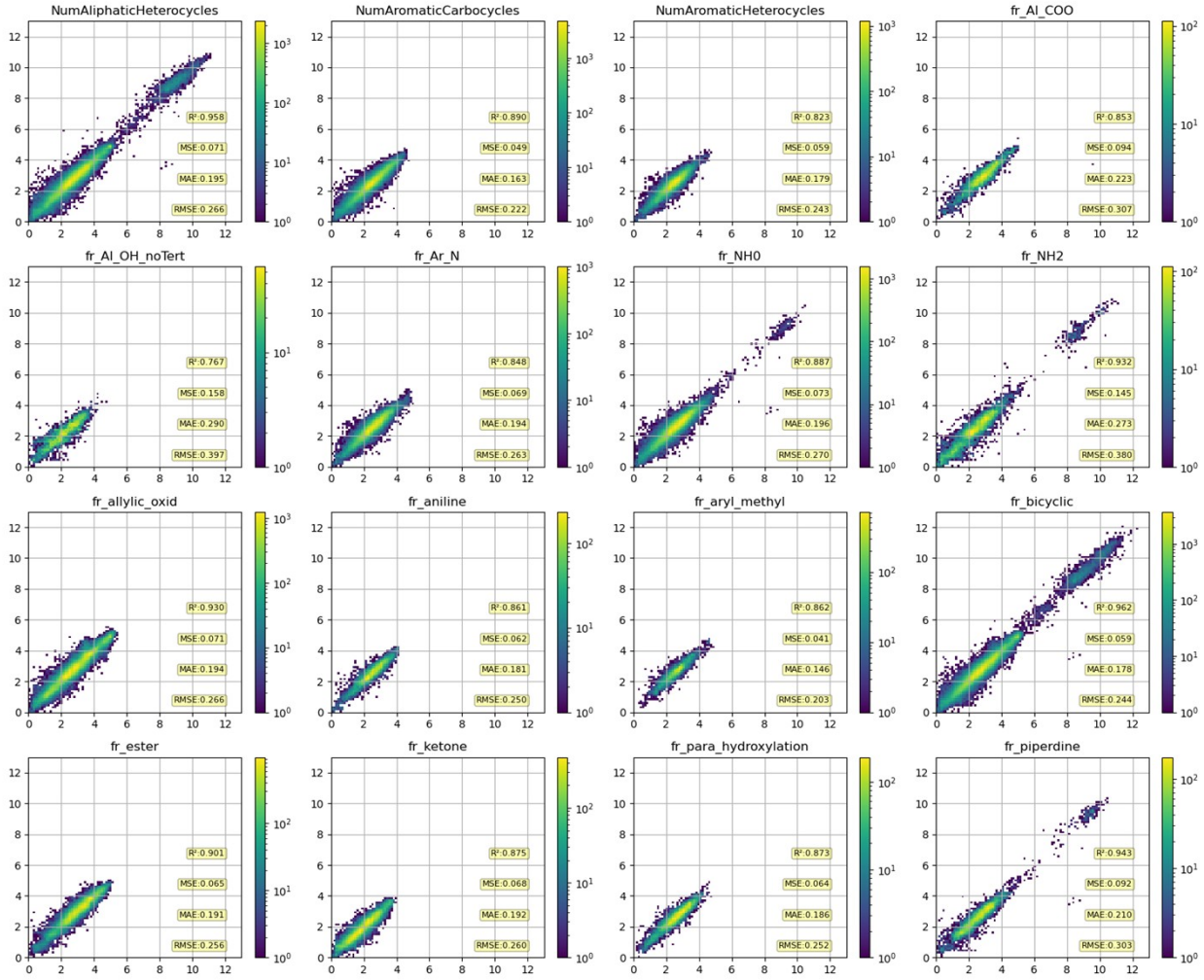


**Figure S2:** The provided scatterplot grid visualizes the relationship between predicted and actual HL-Gap values for various molecular subsets, each defined by a specific molecular descriptor or fragment count. Each plot displays predicted HL-Gap (GAP_P) on the x-axis and actual HL-Gap (GAP) on the y-axis, with color representing the density of molecules in each bin, where warmer colors indicate higher density.

**Verification of Data Split Consistency:**

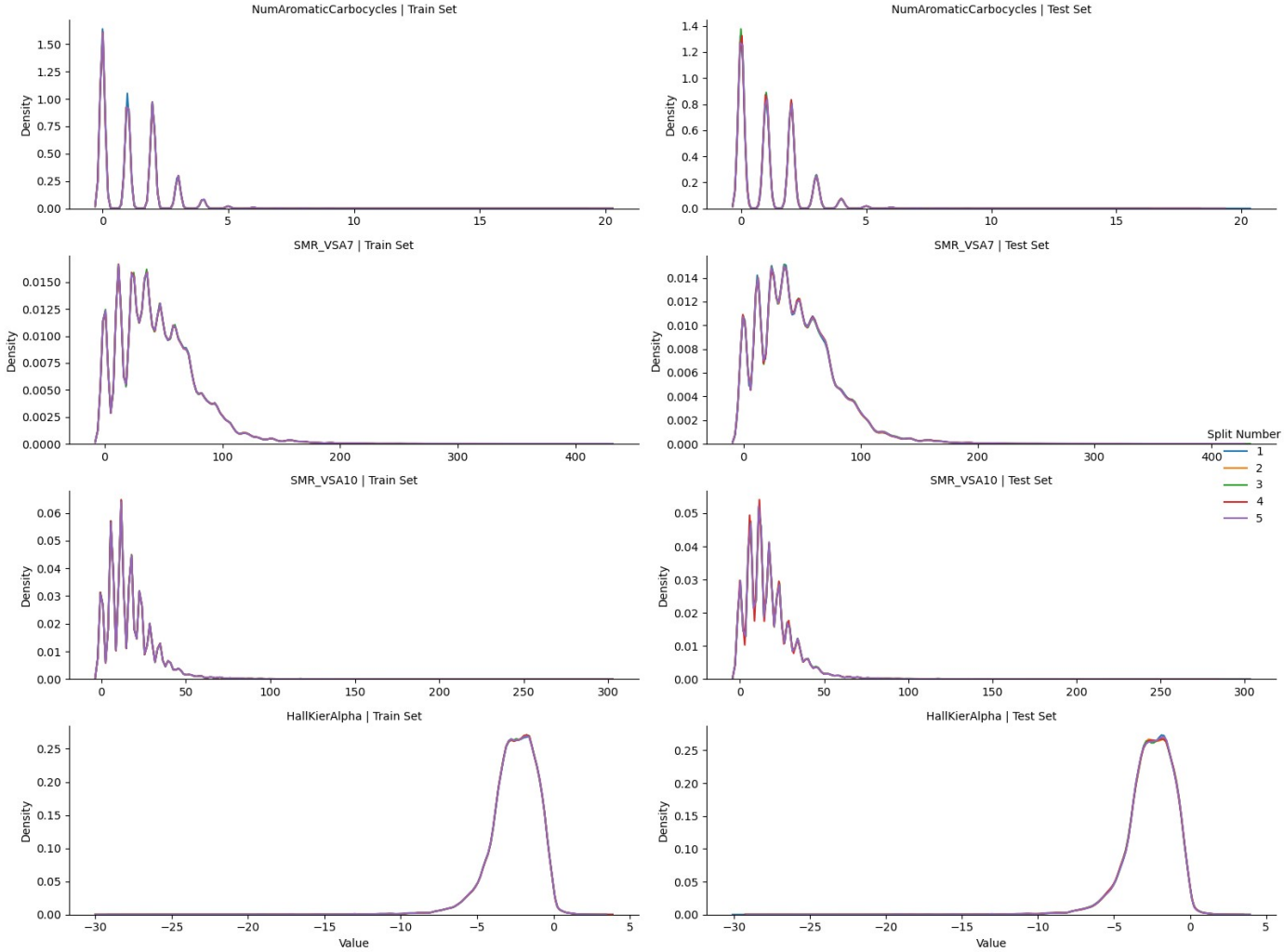


**Figure S3:** This grid of Kernel Density Estimate (KDE) plots visualizes the statistical distribution of four key molecular descriptors. Each row corresponds to a specific descriptor, and each column represents either the training (70%) or testing (30%) subset. Within each subplot, the five overlapping curves represent the data distribution from five independent, random splits. The high degree of overlap between the curves demonstrates that the data distributions are highly consistent and stable, validating that the random splitting methodology is robust and unbiased.