Supplementary Information (SI) for Digital Discovery.

Supporting Information

Unveiling the Impact of Ligand Configurations and Structural Fluxionality on Virtual Screening of Transition-Metal Complexes

Adarsh V. Kalikadien,[†] Niels J. van der Lem,[†] Cecile Valsecchi,[‡] Laurent Lefort,[¶] and Evgeny A. Pidko*[,†]

[†]*Inorganic Systems Engineering, Department of Chemical Engineering, Faculty of Applied Sciences, Delft University of Technology, Van der Maasweg 9, 2629 HZ, Delft, The Netherlands*

[‡]*Discovery, Product Development and Supply, Janssen Cilag S.p.a., Viale Fulvio Testi, 280/6, 20126 Milano, Italy.*

[¶]*Discovery, Product Development and Supply, Janssen Pharmaceutica N.V., Turnhoutseweg 30, 2340 Beerse, Belgium.*

E-mail: e.a.pidko@tudelft.nl

Contents

S1 Substrate binding energy results

In this study, the substrate binding energy of acetonitrile was used as an alternative target property to the stability of a metal-ligand complex.[1]

To obtain substrate binding energies, an additional step to the workflow, outlined in Figure S4, was introduced. Following complex sorting, automated removal of the acetonitrile substrate from the complex structures was performed (*vide infra*). The substrate binding energy is defined by the following formula:

$$E_{bind} = E_{DFT,opt,complex} - (E_{DFT,opt,complex-nosubs} + E_{DFT,opt,subs}) \tag{1}$$

In this equation, substrate binding energy is described by the DFT optimized energy differences between the complex, ($E_{DFT,opt,complex}$) minus the sum of substrate-removed complex ($E_{DFT,opt,complex-nosubs}$) and the energy of non-bonded substrate ($E_{DFT,opt,subs}$). Performed analysis for substrate binding energy is the same as for stability. First, energies between ligand configurations are compared to investigate trends in configuration most favorable in energy.

S1.1 Energy differences between ligand configurations

Binding energy results are visualized in Figure S1a, On the left side, interaction ΔE_{bind} between ligand configurations are plotted for each metal center. Binding energy differences are depicted in the right graphs. Figure S1a illustrates the interaction ΔE_{bind} among iridium ligand configurations. At the top of this figure, the general reference structures of the ligand configurations are displayed, whereas the structures for the other ligand configurations are depicted at the bottom of the graph. In this figure, energy of the H-N axial ligand pair complexes are compared with with H-H axial ligand pair complexes for the tested bidentate ligands. A negative ΔE_{bind} is observed for the majority of bidentate ligands. This indicates that energy of the H-H axial ligand pair configuration is more favorable in energy over the

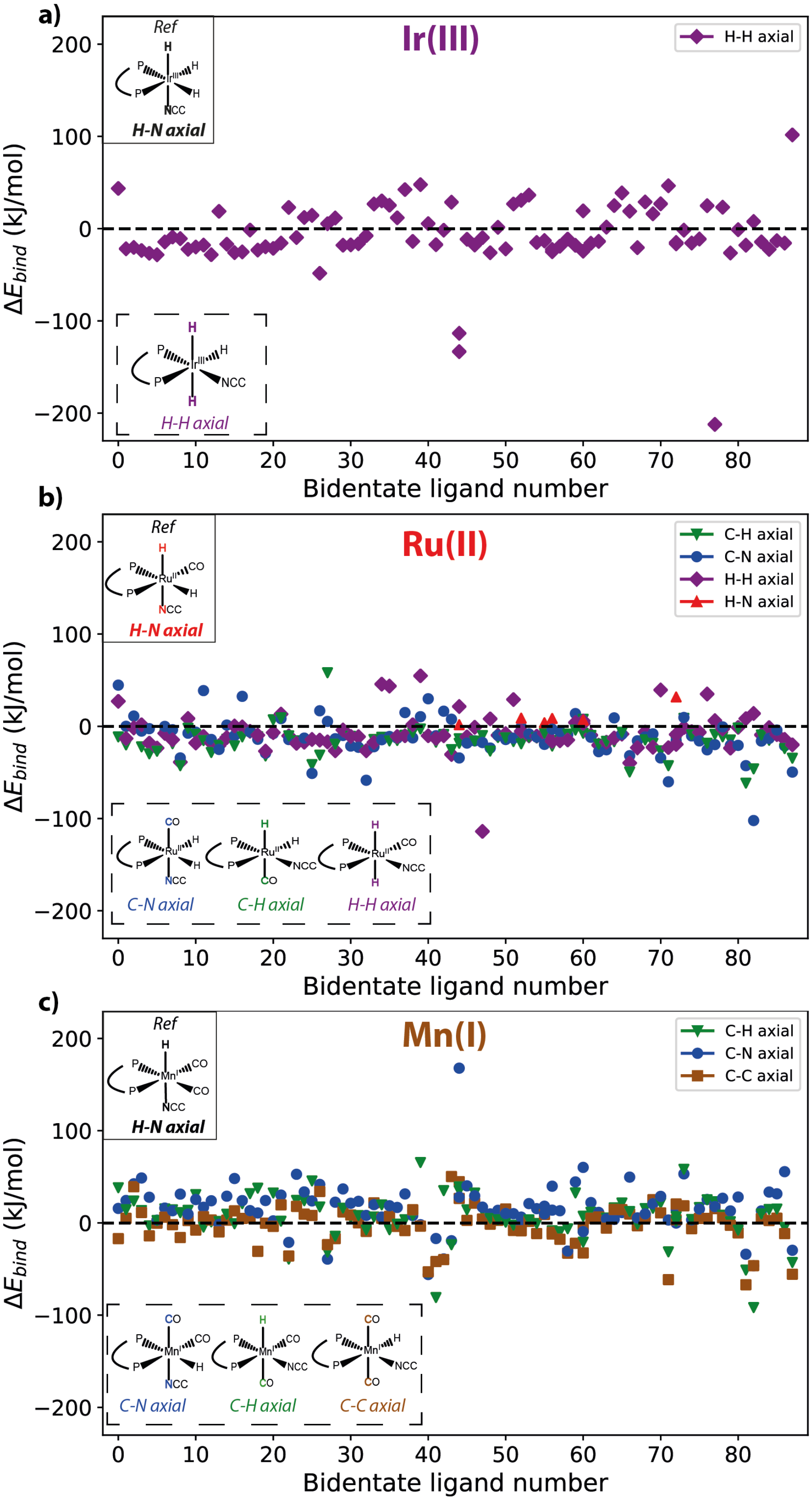

Figure S1: Substrate binding energy differences between a reference configurational isomer, shown at the top of the graphs, and other configurational ligand configurations, shown in the bottom of the graphs, for set of bidentate ligands for iridium (a), ruthenium (b) and manganese (c) complexes

H-N axial ligand pair ligand configurations for most bidentate ligands.

Figure S2b shows the binding energy differences between ligand configurations of ruthenium complexes. In this figure, the energy of complexes with axial ligands C-H, C-N and H-H are compared with that of complexes featuring the H-N axial ligand pair. General structure for the reference H-N axial ligand pair is shown at the top of the figure, and compared structures are shown at the bottom. The lowest energy configurational configuration often varies among the different bidentate ligands. However, a positive ΔE_{bind} is observed for a majority of ligand configurations, indicating that the H-N axial configuration is clearly not favorable in terms of substrate binding energy in comparison to the other ligand configurations.

ΔE_{bind} data for manganese ligand configurations is shown in Figure S1c. In this figure, substrate binding energies of the C-N, C-H and C-C ligand configurations are compared to the substrate binding energy of the H-N isomer. A significant number of ligand configurations show a positive ΔE_{bind}. However, no clear majority of a configuration at the global minimum in energy is observed.

Figure S2 illustrates the distribution of lowest binding energy ligand configurations among the axial ligand pairs. However, the preference of iridium complexes is less overwhelming for binding energy as opposed to interaction energy. For binding energies, the H-H axial ligand pair is lowest in energy for 64.7% of bidentate ligands, which is significantly lower than the 92% observed for interaction energies. For ruthenium and manganese complexes, most observed global minimum configurations also changed for binding energies compared to interaction energies. Nevertheless, the determination of the lowest energy configurational configuration varies among the tested bidentate ligands. In most cases, no specific axial ligand pair is significantly more abundantly favorable over the other ligand configurations.

S1.2 Effect of metastable ligand configurations

Figure S3 visualizes configuration statistics for binding energy data of the generated ligand configurations. Figure S3a shows the number of ligand configurations present within 10

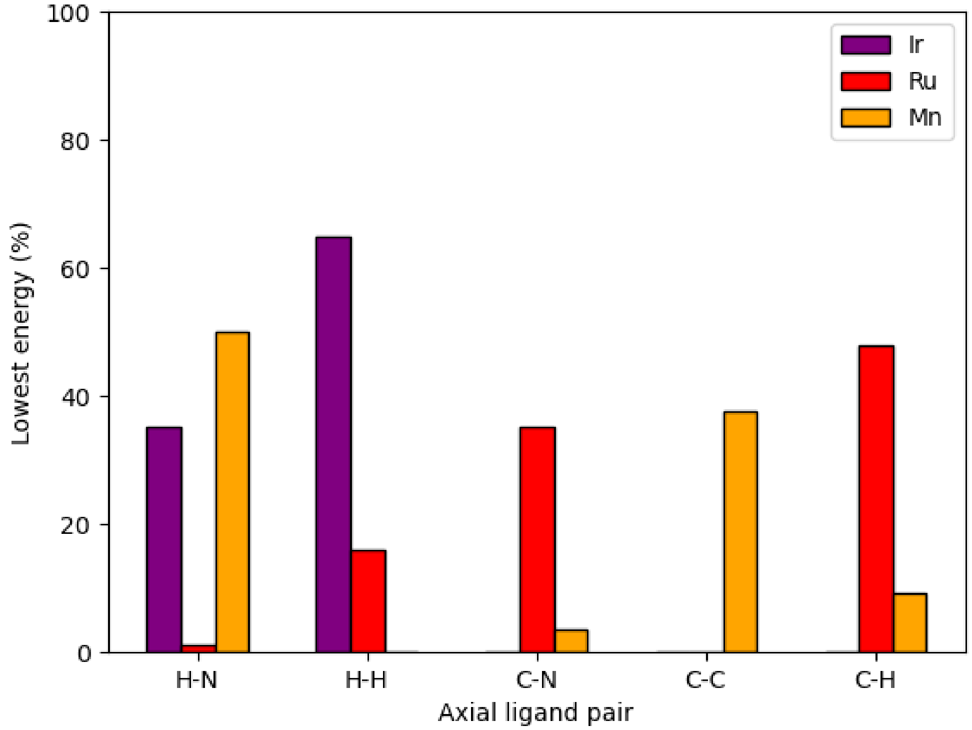

Figure S2: Distribution of the lowest binding energy configuration among axial ligand pairs for the 88 bidentate ligands in combination with 3 metal centers

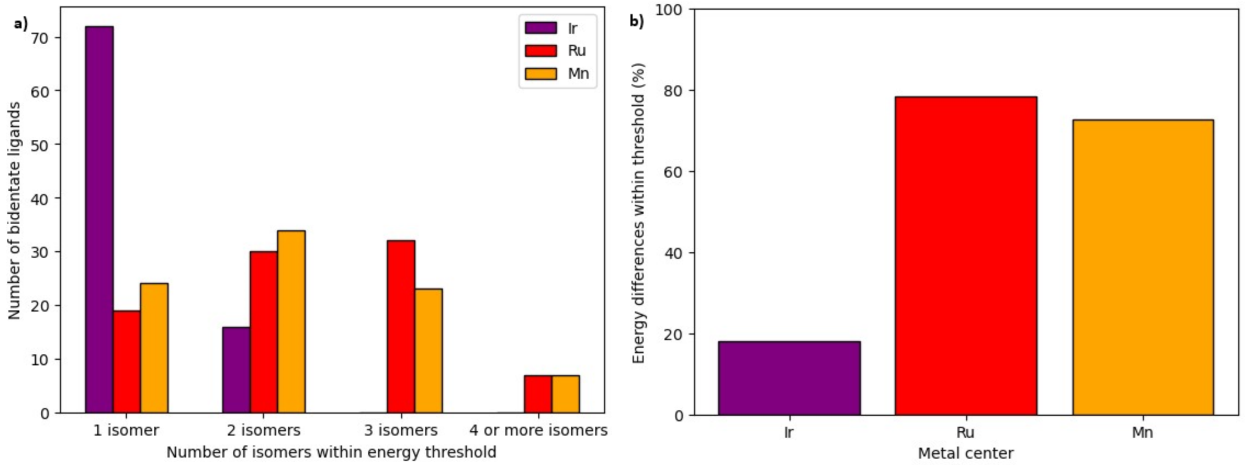

Figure S3: Number of ligand configurations present within a 10 kJ/mol energy range from the lowest binding energy configuration for the researched bidentate ligands (a) and the percentage of bidentate ligands for which multiple ligand configurations are found within the 10 kJ/mol energy range (b) kJ/mol range from lowest binding energy for each bidentate ligand. In Figure S3b, the percentage of bidentate ligands for which multiple ligand configurations have binding energies within 10 kJ/mol from the minimum is visualised. Similarly to interaction energy, 18.2% of iridium complexes contain multiple ligand configurations present within the energy range. Moreover, a slight increase in the percentage of bidentate ligands containing multiple ligand configurations within the 10 kJ/mol energy range is observed. Namely, 78.4% of ruthenium complexes and 72.7% of manganese complexes.

S2 Automated substrate removal

Structures of TM-complexes without the substrate are essential for determining substrate binding energy. The automation of substrate removal has the potential to save time and reduce errors in structures. The general workflow of the substrate remover is shown in Figure S4

The substrate is consistently found at a fixed location for MACE generated xyz files.

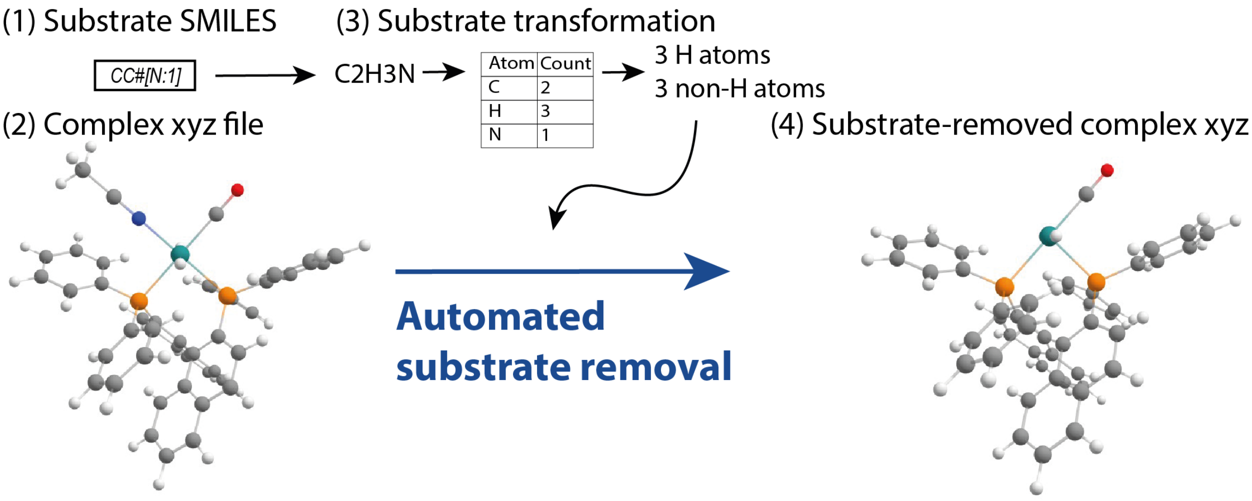

Figure S4: Automated substrate removal workflow, showing inputes (1) and (2), substrate transformation (3) and workflow output (4)

A MACE generated .xyz file consists of individual rows for each atom present in the TM-complex. All hydrogen atoms present are located at the bottom rows of the file. Substrate hydrogen atoms are consistently found at the initial hydrogen rows, while all non-hydrogen substrate atoms are found at the foremost rows. This arrangement of atoms makes substrate removal straightforward if the number of hydrogen and non-hydrogen atoms of the substrate is known.

The following input has to be provided to the substrate remover: TM-complex directory, bidentate ligand number and substrate information. Substrate SMILES have already been supplied for structure generation input. For easy and integrated use of the substrate remover, these SMILES should also be used as input for the substrate remover. However, SMILES only represent the non-hydrogen atoms of the substrate. Fortunately, SMILES can be converted into chemical formulas through RDKit tools, as shown in Figure S4-3. With the chemical formula known, it becomes feasible to iterate over this formula to obtain the specific types and quantities of atoms present in the substrate. From this information, the number of hydrogen and non-hydrogen atoms can be determined.

Once substrate information is known, the substrate is removed from the MACE generated TM-complex. A new directory called *substrate_removed* is created to save the newly created complexes without substrate. The suffix *no_substrate* is added to the end of each individual file to denote the difference from the original file.

S3 Details on machine learning

Prior to applying machine learning (ML) algorithms, the dataset was pre-processed to ensure consistency and relevance. Specifically, only the lowest-energy conformers of the metal-ligand complexes were selected, ensuring that each configuration was represented by a single vector of physical-chemical descriptors. In the OBeLiX package, donor atoms in the bidentate ligand are labeled as either ”min” or ”max.” To standardize the descriptor representation for ML and maintain consistency in donor atom labeling, the H-N configuration was chosen as a reference. This ensured that descriptors dependent on the ”min” and ”max” labels were oriented uniformly across all configurations.

The Random Forest (RF) model was implemented in our ML pipeline which can be found on our Github page (https://github.com/epics-group/obelix-ml-pipeline). Code from this pipeline was modified to allow modeling with Logistic Regression (LR). The scikit-learn Python package was used for all functionalities included this pipeline. For RF, feature importances were calculated using the deault Gini importance as implemented in SKlearn. The feature importances from LR were calculated by multiplying the magnitude of coefficients with the standard deviations of the corresponding parameters in the data.

An 80/20 train/test split was used for the in-domain modeling approaches. For out-of-domain modeling approaches, a fixed set of 16 ligands and their configurations were kept out of the training set. For each training, a 5-fold cross-validation method is applied. A grid search cross-validation method was used for selecting hyperparameters. Within this grid search, the options for each hyperparameter of the RF models were:

- ’bootstrap’ = [False],
- ’max_depth’ = [5, 50, 100, None], *None applied only to OHE-based models
- ’max_features’ = [3, 5],
- ’min_samples_leaf’ = [1, 2, 5, 10],
- ’min_samples_split’ = [2, 5, 10],
- ’n_estimators’ = [50, 100, 200],

The options for each hyperparameter of the LR models were:

- ’penalty’ = [’l1’], ’l2’, ’elasticnet’, None],
- ’C’ = [0.001, 0.01, 0.1, 1, 10,100],
- ’solver’ = [’liblinear’, ’lbfgs’]

For modeling the stability of ligand configurations, the class_weight variable was added with the following options: 'class_weight' = ['balanced', 'balanced_subsample', None].

S4 Transferability of descriptors across ligand configurations and metal centers

Figure S5 shows a heatmap for all calculated descriptors and the resulting R^2. Each heatmap shows all possible 'metal, configuration' combinations on the x- and y-axes.

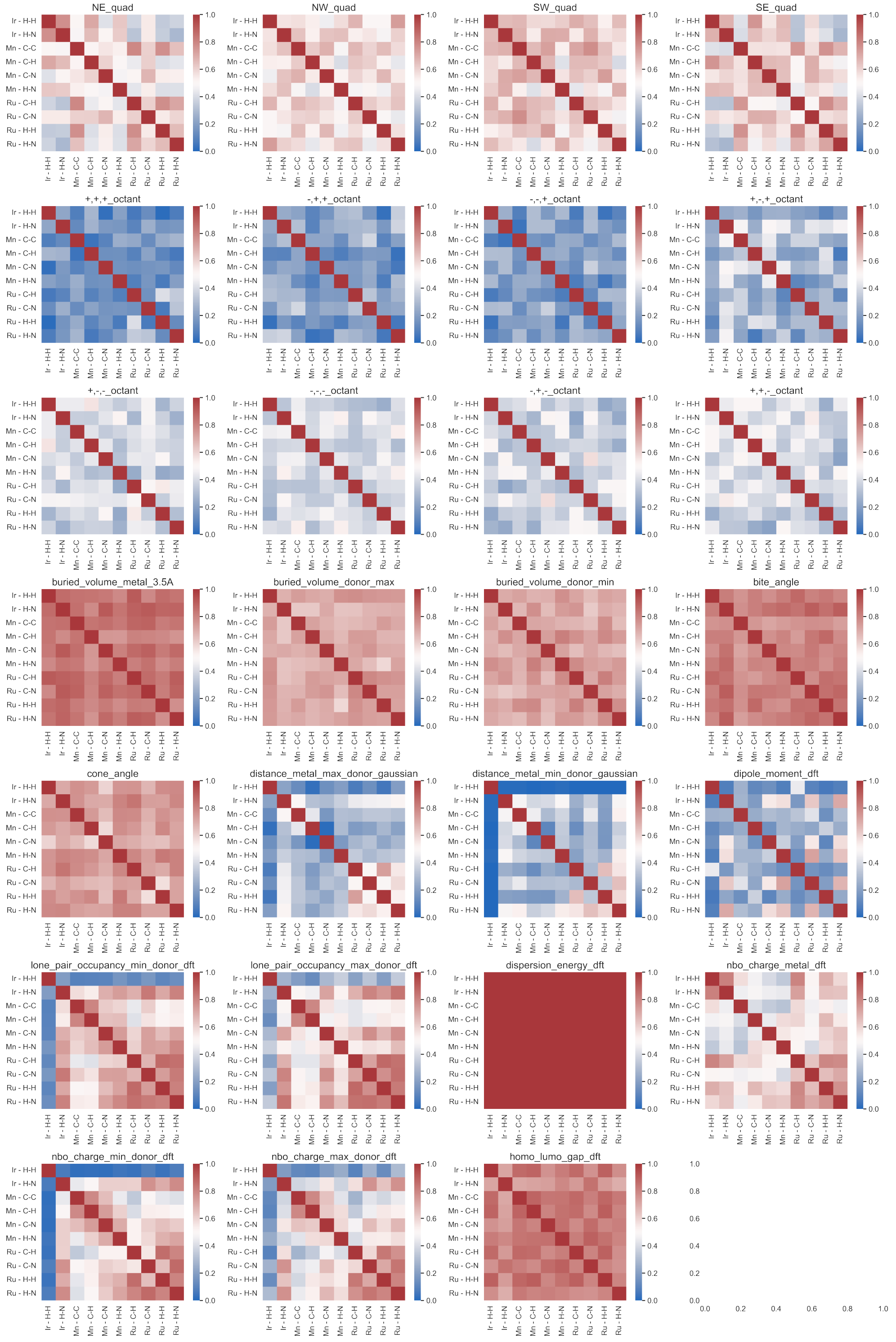

Figure S5: Matrices for all calculated descriptors with R^2 scores of linear models between specific descriptor from one set of bidentate ligands with a specific metal and ligand configuration to another set with a different combination of metal and ligand configuration.

S5 Orbital visualization

In this section, visualized orbitals for two bisphosphine bidentate ligands are shown. Figure S6 shows the HOMO and LUMO orbitals for iridium complexes, Figure S7 shows the HOMO and LUMO orbitals for ruthenium complexes, and Figure S8 shows the HOMO and LUMO orbitals for manganese complexes. The HOMO, LUMO and HOMO-LUMO gap energies are widely used electronic descriptors, as they can be related to the complex reactivity and stability. Firstly, the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) of the different complex configurations are visualized to illustrate the differences in the electronic structure. The figures reveal a similar LUMO shape among the different ligand configurations, primarily localized on the bidentate ligand backbone. However, substantial variations in the nature of HOMO can be observed for different ligand configurations.

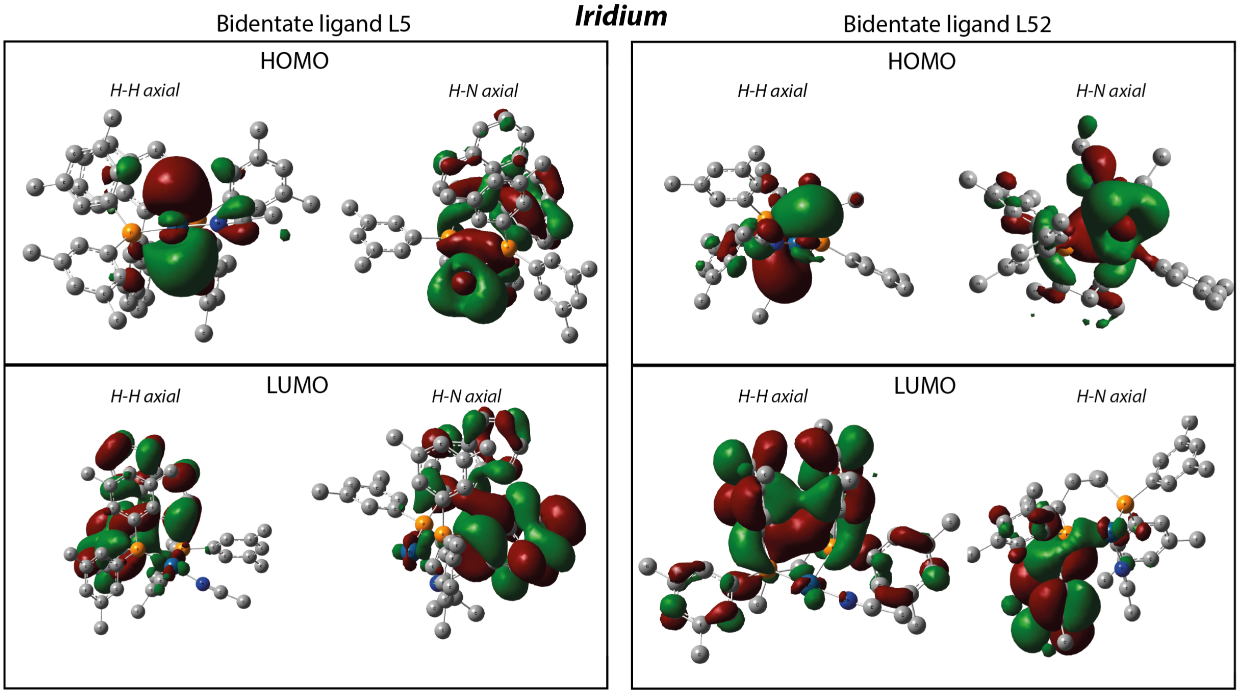

Figure S6: HOMO and LUMO orbital visualizations for different ligand configurations for iridium complexes

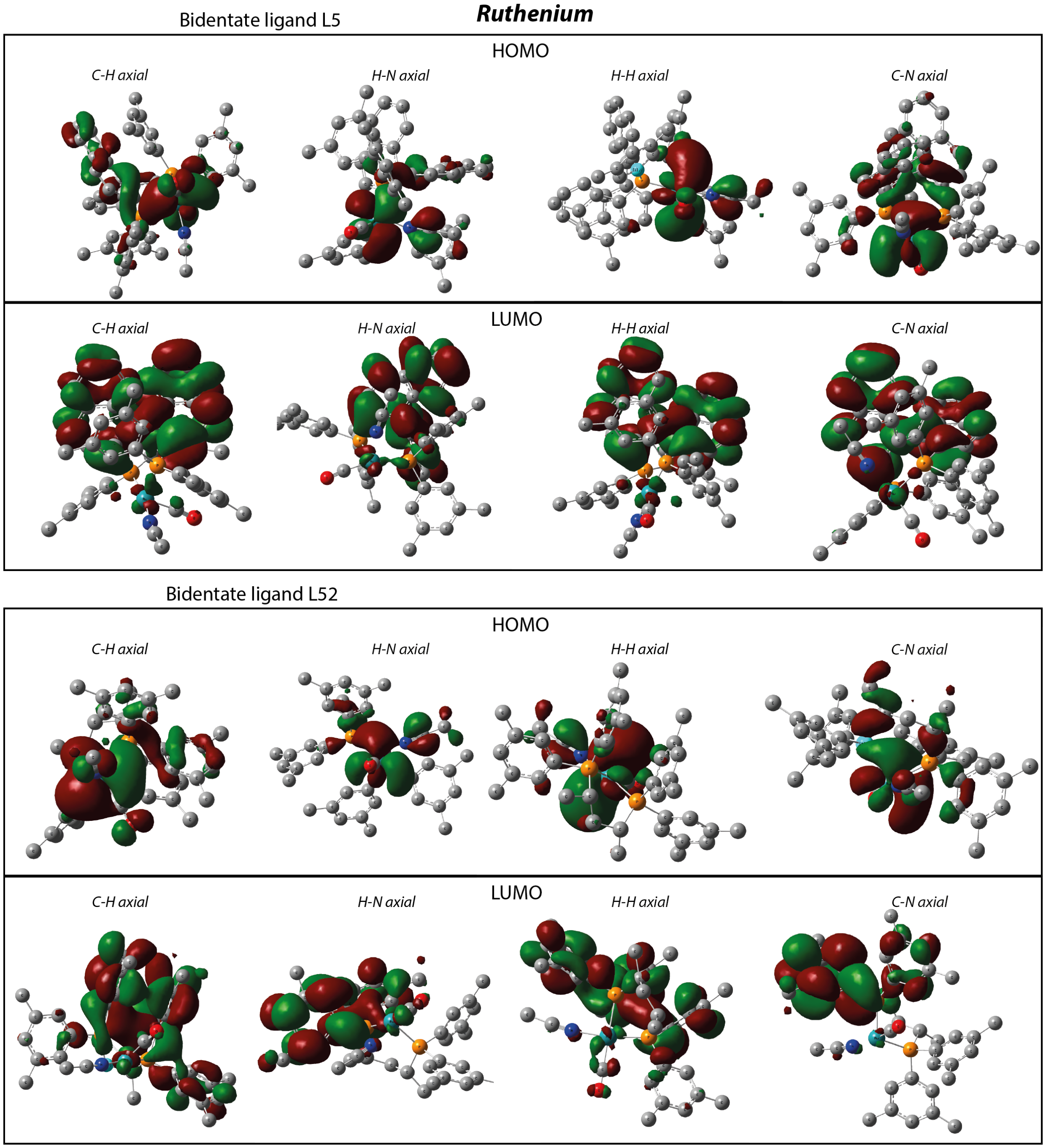

Figure S7: HOMO and LUMO orbital visualizations for different ligand configurations for ruthenium complexes

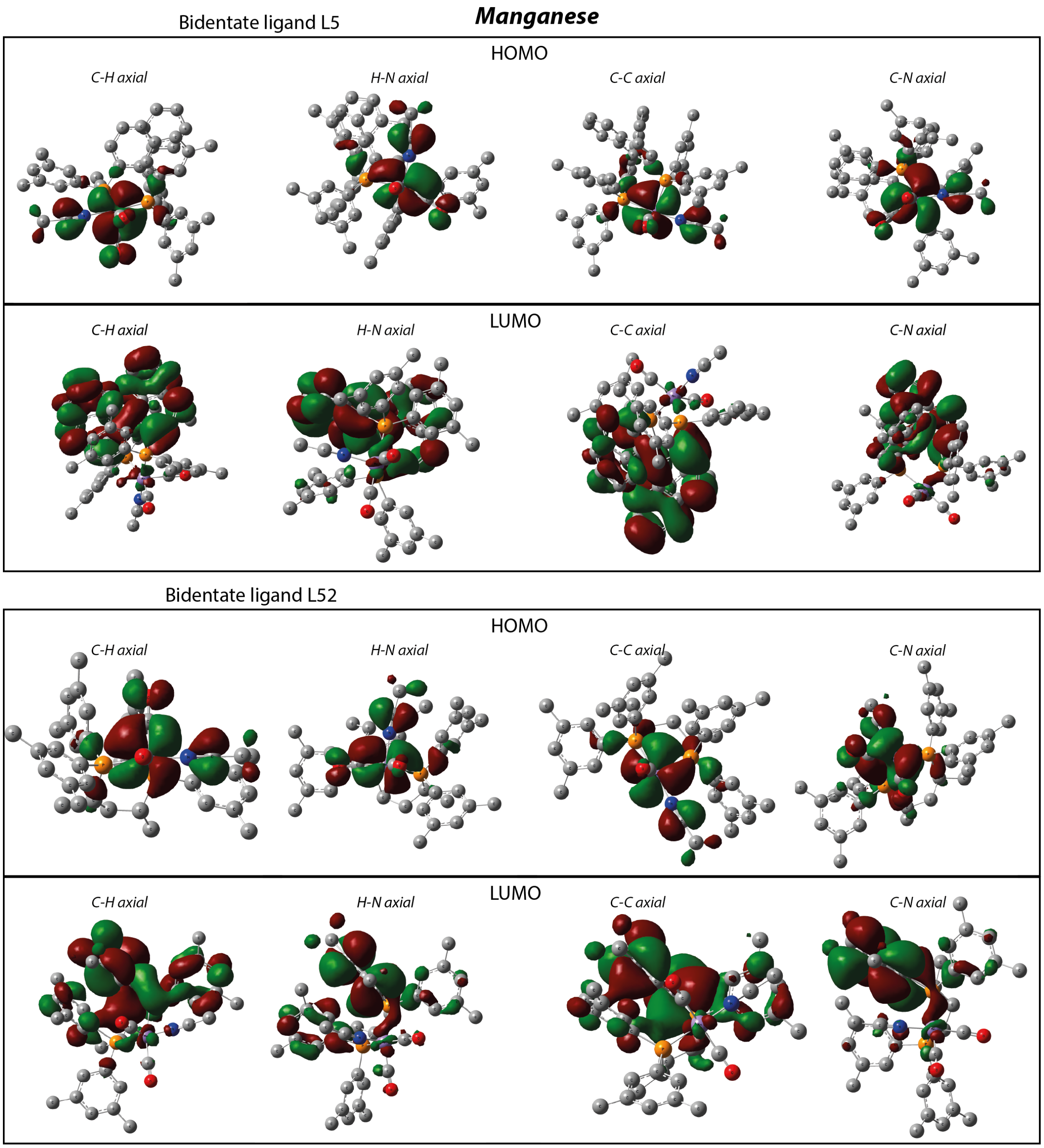

Figure S8: HOMO and LUMO orbital visualizations for different ligand configurations for manganese complexes

S6 In-domain modeling of axial ligand configuration pair

S6.1 Feature importances

The feature importances of discussed ML models are reported in Figure S9 for modeling on the dataset with all metal centers and Figure S10 for modeling on a subset of the data with only Mn(I) metal centers.

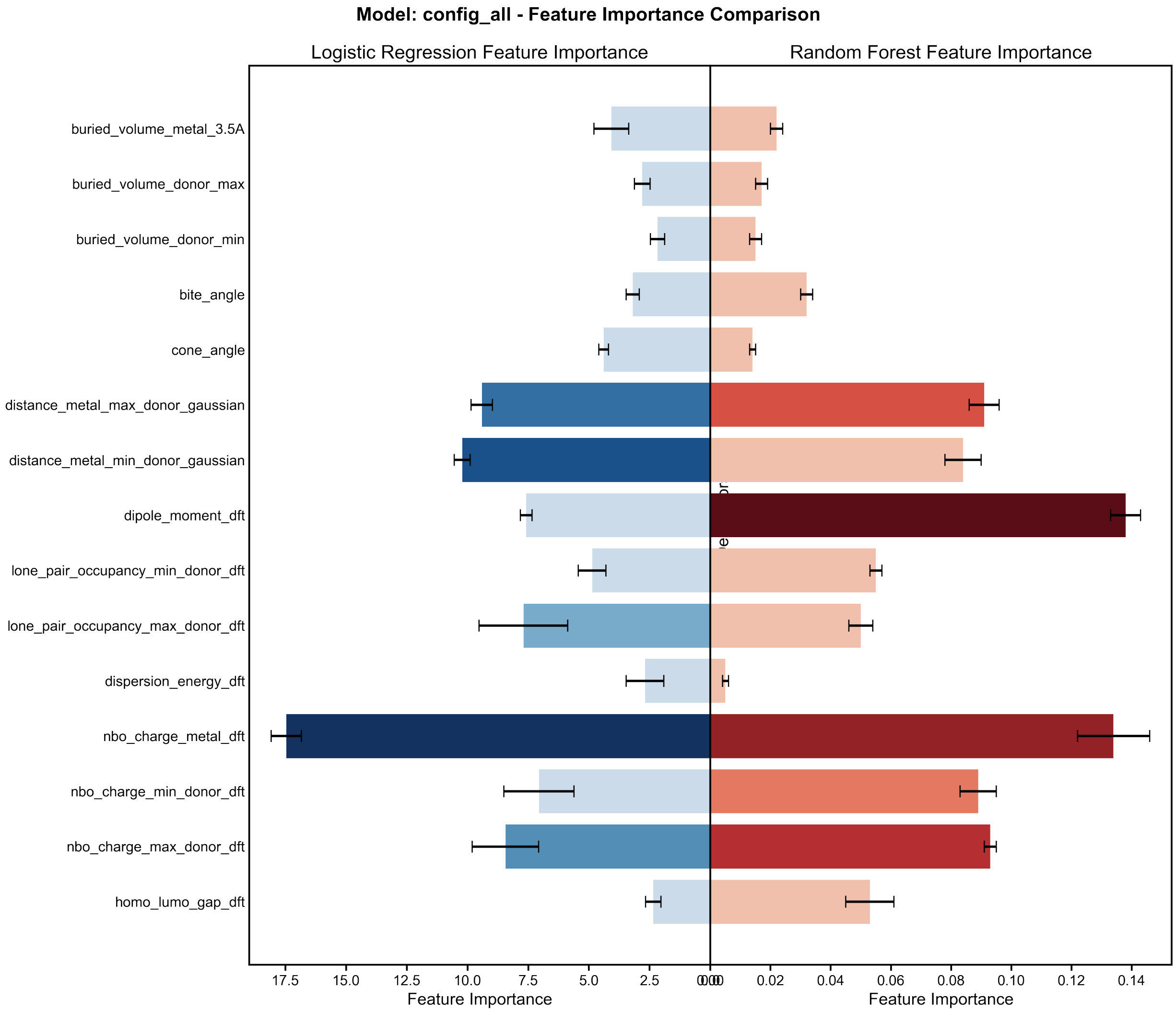

Figure S9: Feature importance of ML modeling of configuration over dataset with all metals and ligand configurations

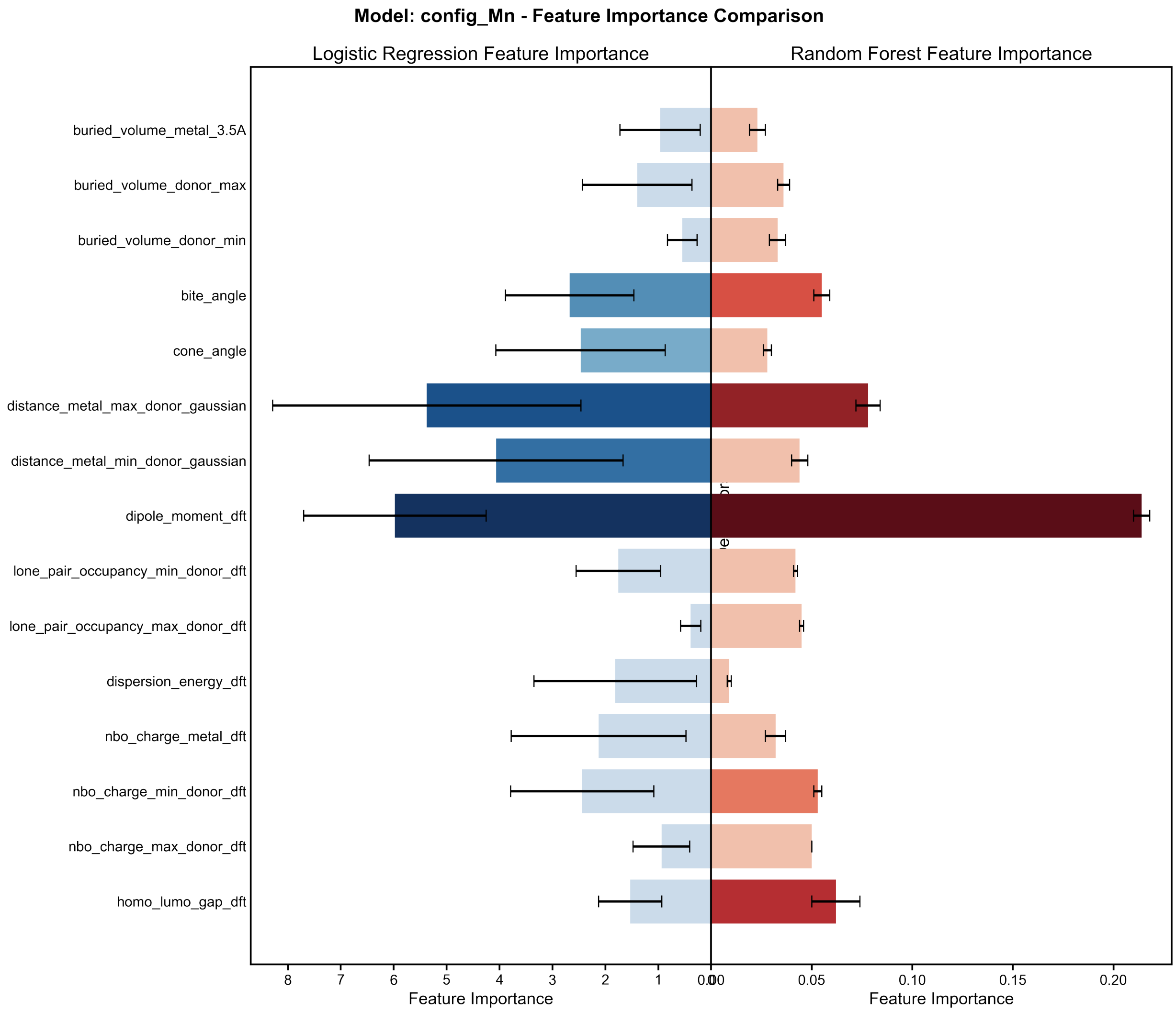

Figure S10: Feature importance of ML modeling of configuration over subset of data with Mn as the metal center

S7 Out-of-domain modeling of axial ligand configuration pair

S7.1 ML performance

For ML modeling to distinguish between different axial configurations, we tested an out-of-domain approach in which 16 ligands were kept as the test set, simulating a case of applying the trained models to fully new ligands. The results are reported in Figure S13 and show good results on the full dataset or a subset with Ru(II) and Ir(III) metal centers. However, similarly to the in-domain modeling, a performance drop is observed for the subset with Mn(I) metal centers.

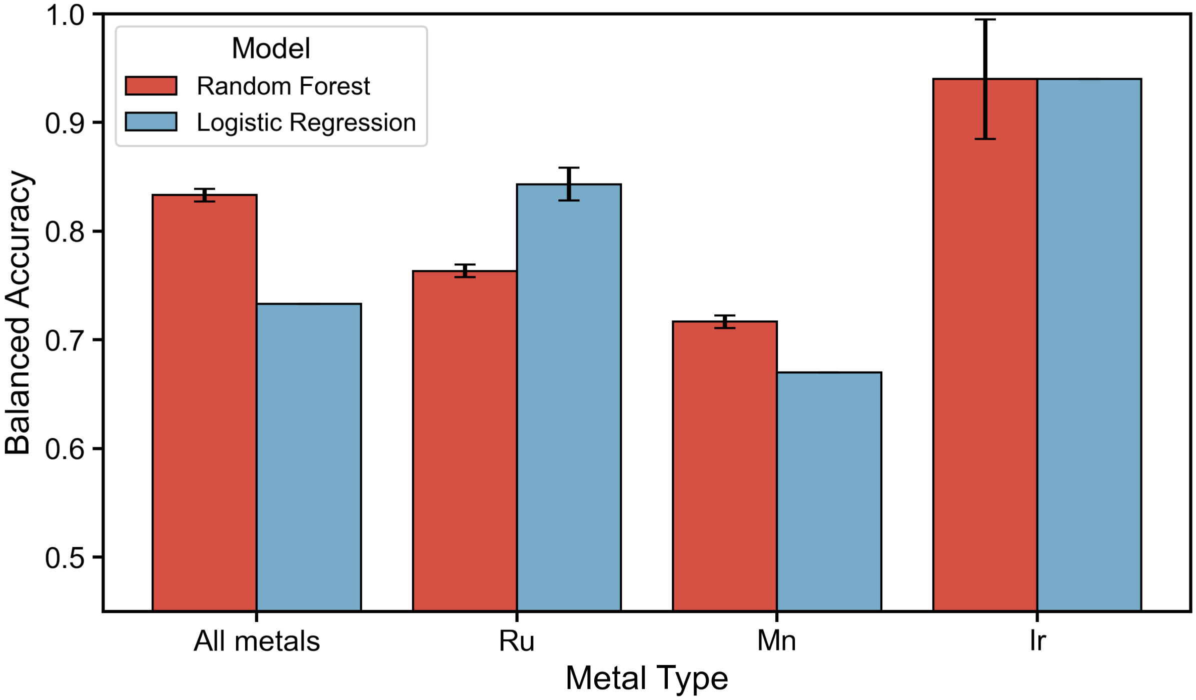

Figure S11: Performance metrics for the out-of-domain modeling of ligand configurations. The performance of RF and LR are displayed in a red and blue bar respectively. The y-axis denotes the Balanced Accuracy score and the x-axis specifies whether modeling is done on the dataset containing all metal centers and ligand configurations or on a metal-specific subset.

S8 In-domain modeling of stability of ligand configurations

S8.1 Feature importances

The feature importances of discussed ML models are reported in Figure S12 for modeling on a subset of the data with only Ir(III) metal centers.

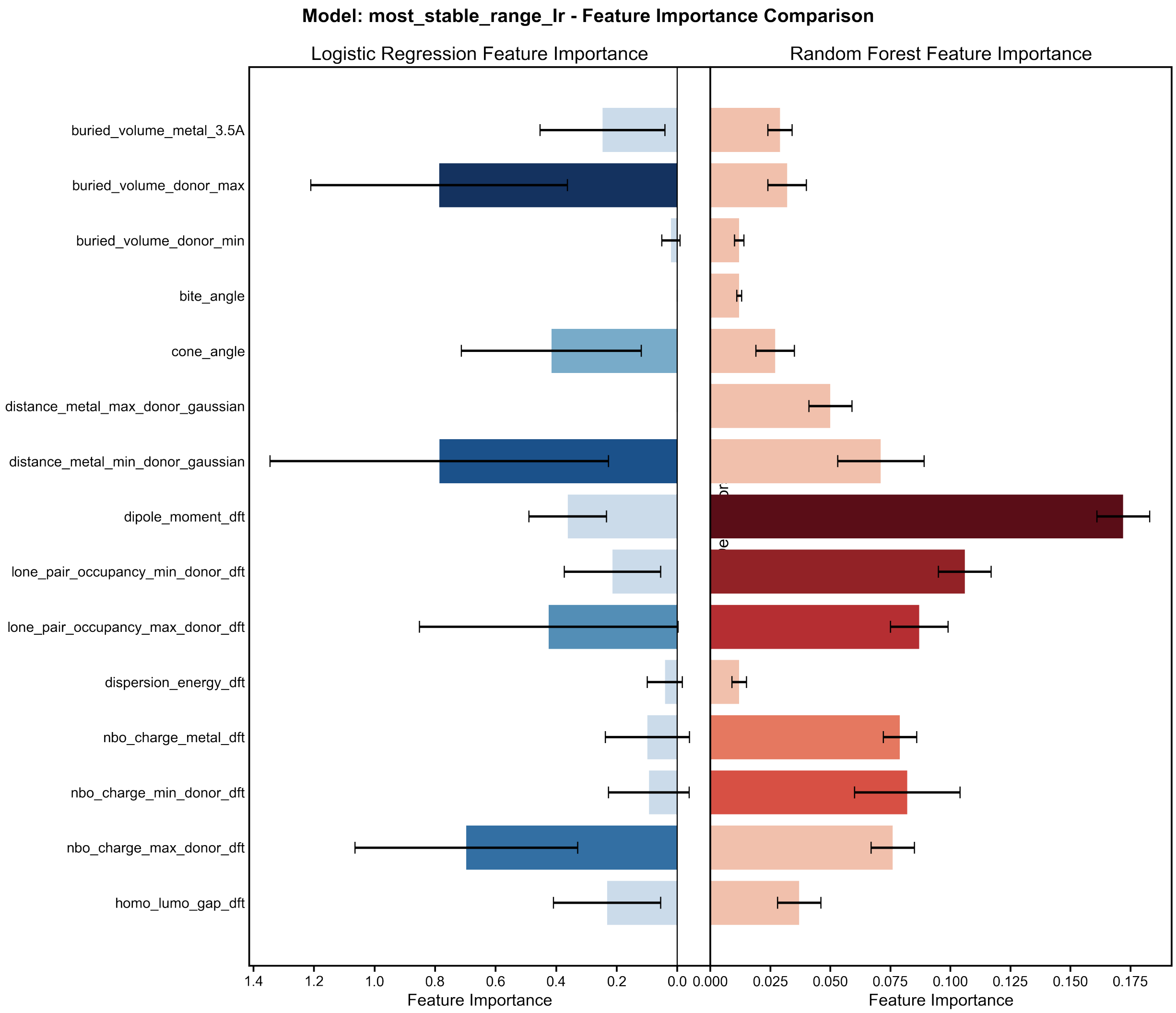

Figure S12: Feature importance of ML modeling of stability of configurations over dataset with Ir and ligand configurations

S9 Out-of-domain modeling of stability of ligand configurations

S9.1 ML performance

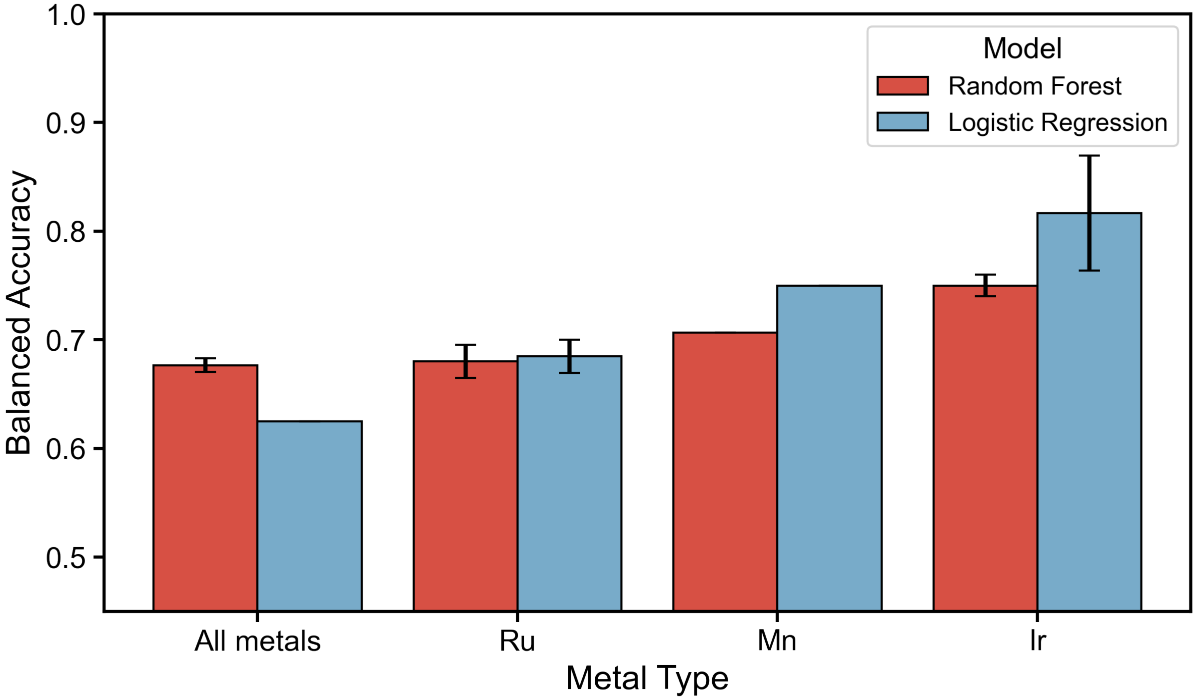

Figure S13: Performance metrics for the out-of-domain modeling of the stability of ligand configurations. The performance of RF and LR are displayed in a red and blue bar respectively. The y-axis denotes the Balanced Accuracy score and the x-axis specifies whether modeling is done on the dataset containing all metal centers and ligand configurations or on a metal-specific subset.

References

(1) Meyer, B.; Sawatlon, B.; Heinen, S.; von Lilienfeld, O. A.; Corminboeuf, C. Machine learning meets volcano plots: computational discovery of cross-coupling catalysts. *Chem. Sci.* **2018**, *9*, 7069–7077.