Alkaline Earth Metal Binding with N-Heterocyclic Carbenes: Machine Learning-Assisted Screening of Metal-Ligand Interactions

Saurabh Singh Negi,[a] Puneet Gupta*[a,b]
[a] Computational Catalysis Center, Department of Chemistry, Indian Institute of Technology, Roorkee, Uttarakhand - 247667,
[b] Center for Sustainable Energy, Department of Chemistry, Indian Institute of Technology, Roorkee, Uttarakhand- 247667, India,
India. E-mail: puneet.gupta@cy.iitr.ac.in

Contents

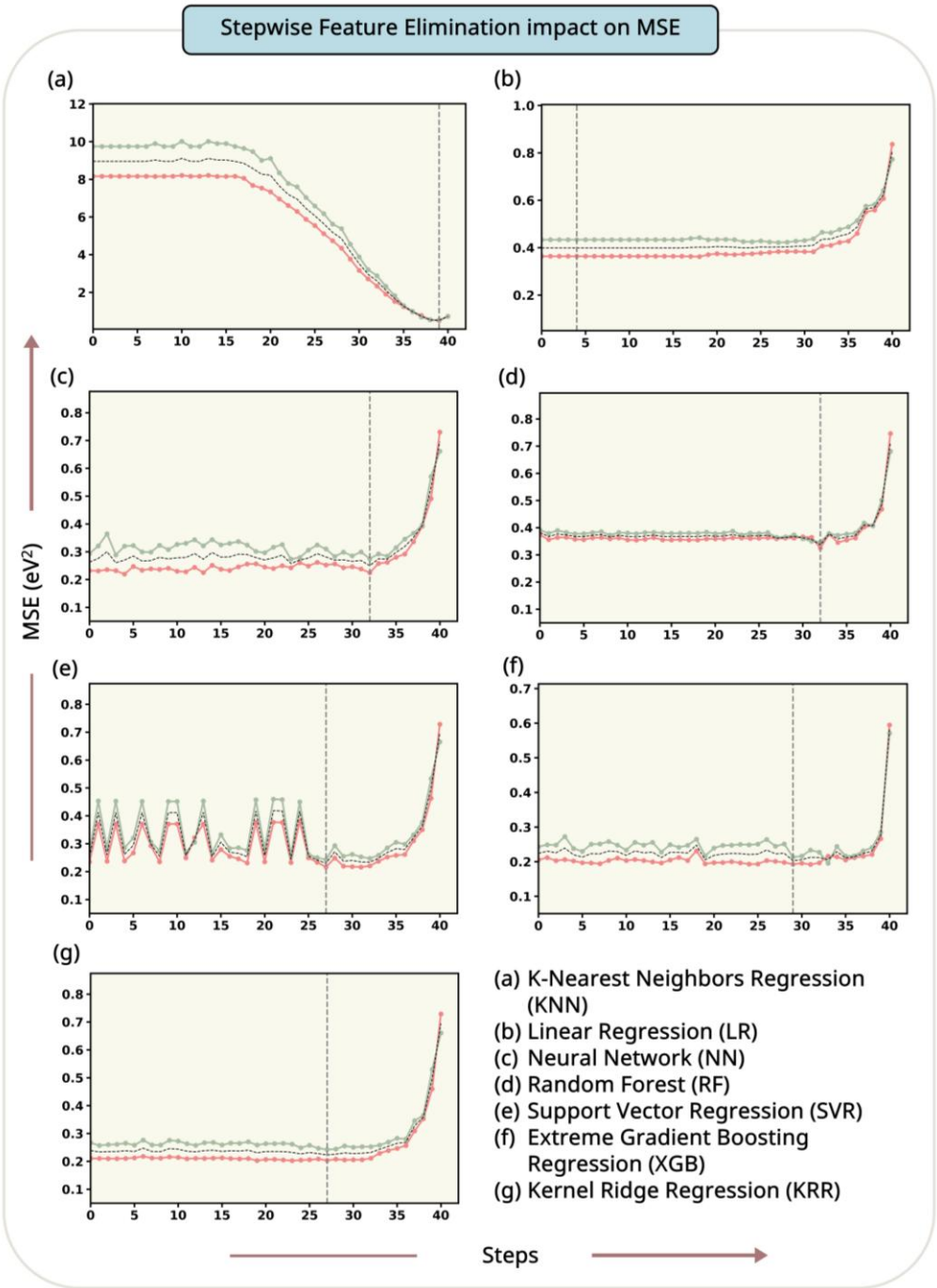

Fig S1.1. Change in mean square error (MSE) values using **RDKit descriptors** for different models with reducing the number of descriptors. The red line is training MSE and green line is testing MSE, vertical grey line is marking the step corresponding to min(train + test) MSE value.

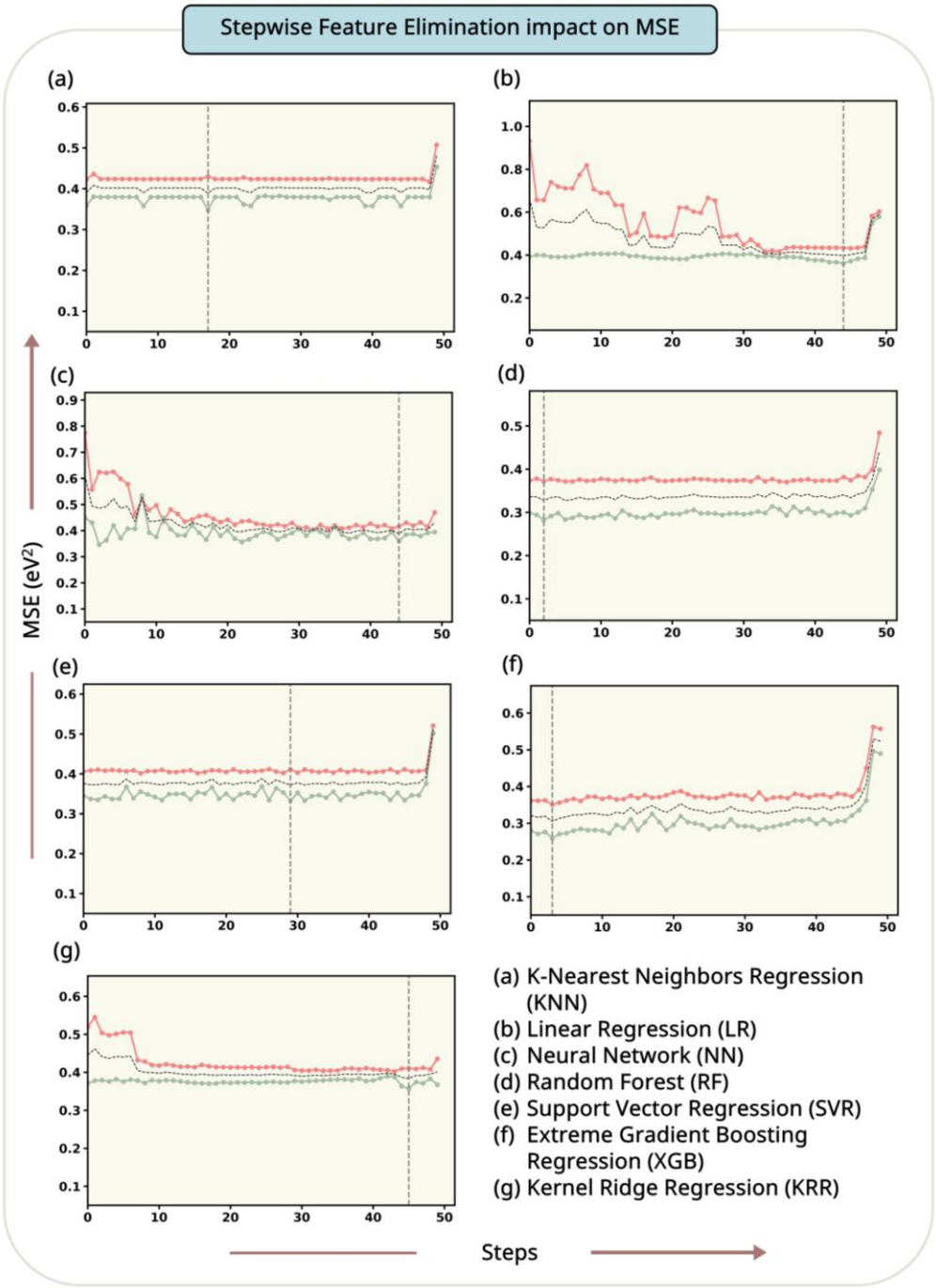

Fig S1.2. Change in mean square error (MSE) values using **Sterimol descriptors** for different models with reducing the number of descriptors. The red line is training MSE and green line is testing MSE, vertical grey line is marking the step corresponding to min(train + test) MSE value.

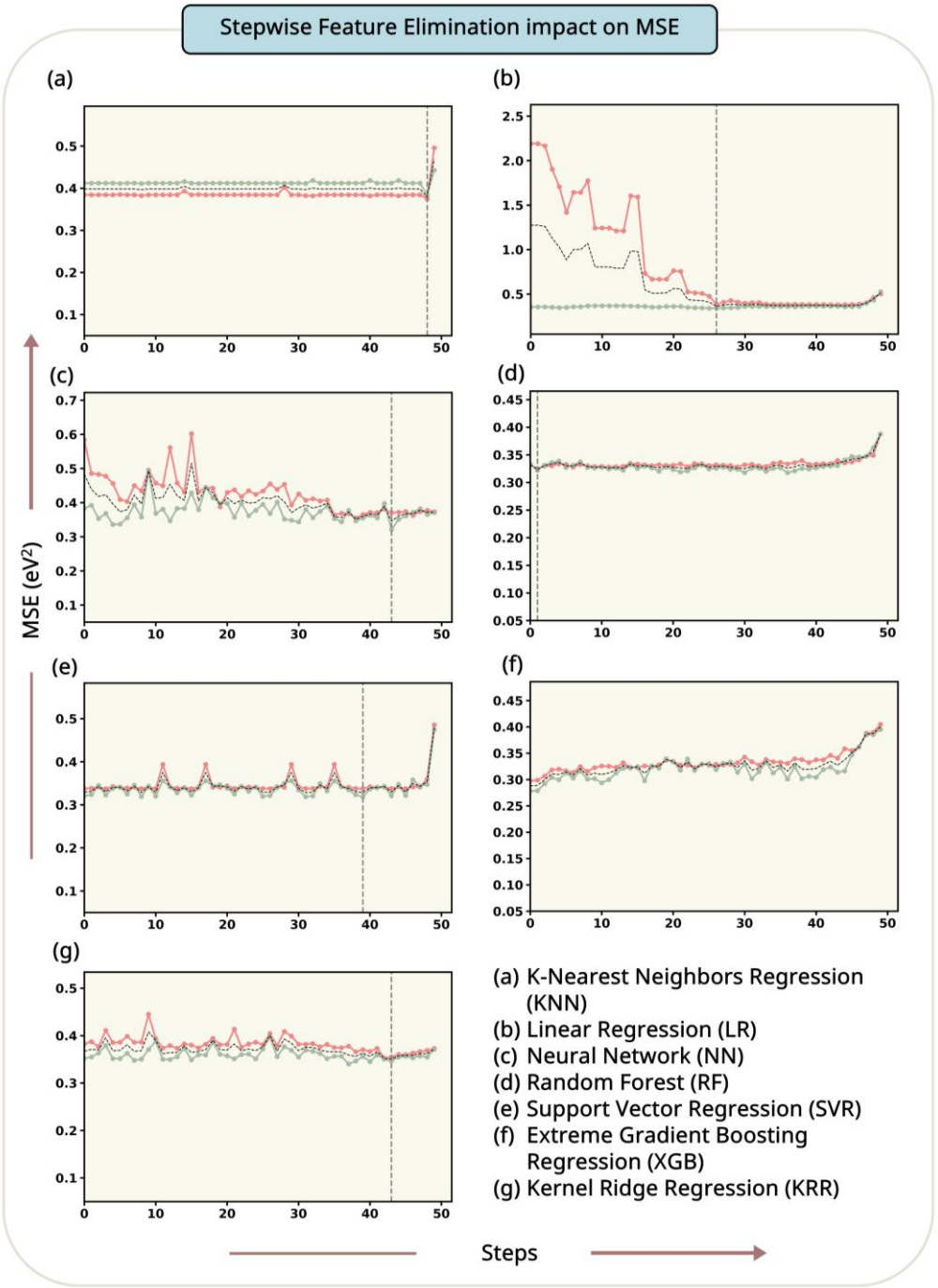

Fig S1.3. Change in mean square error (MSE) values using **Sterimol descriptors without Be complexes** for different models with reducing the number of descriptors. The red line is training MSE and green line is testing MSE, vertical grey line is marking the step corresponding to min(train + test) MSE value.

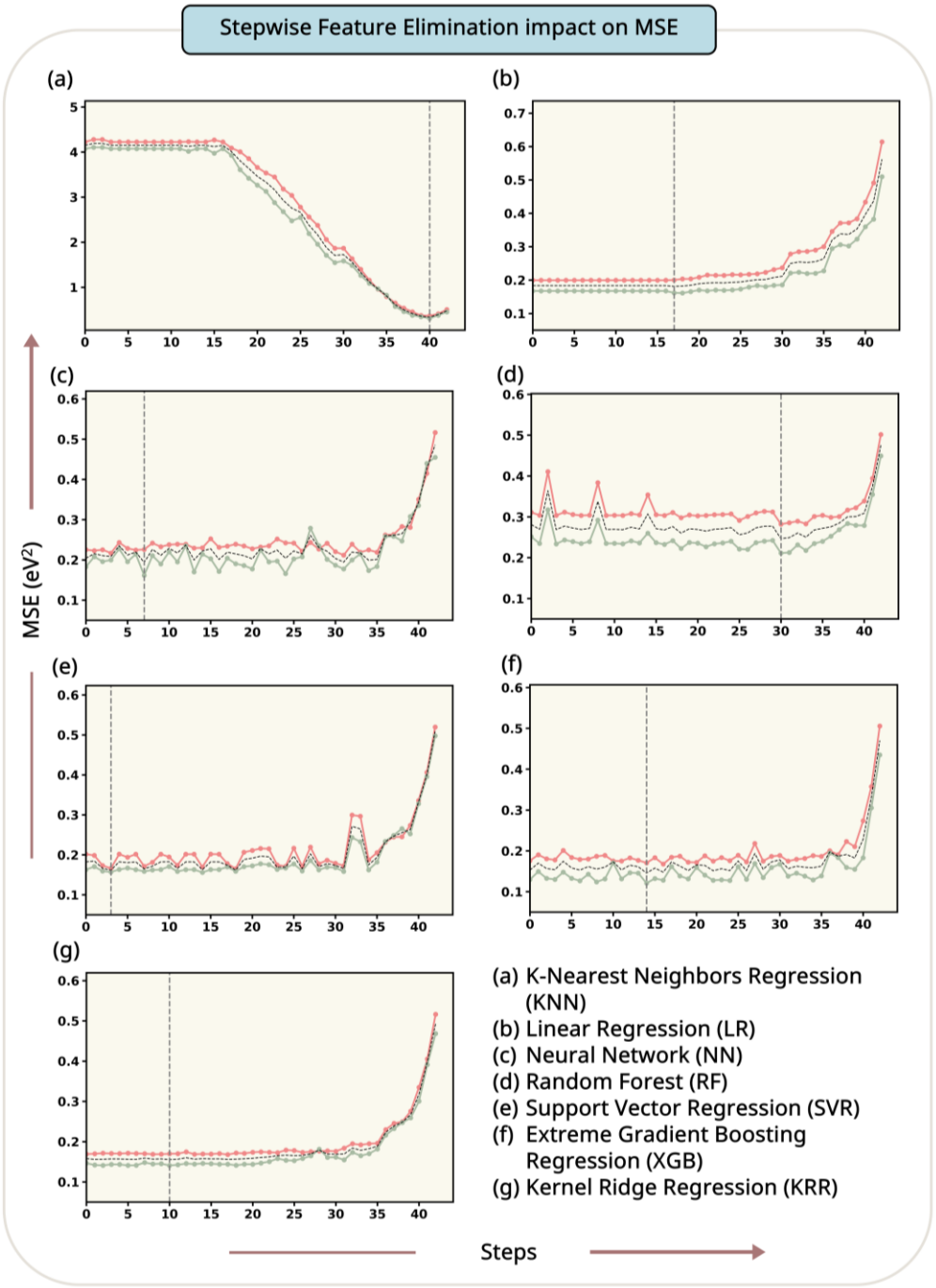

Fig S1.4. Change in mean square error (MSE) values using **RDKit + Sterimol descriptors** for different models with reducing the number of descriptors. The red line is training MSE and green line is testing MSE, vertical grey line is marking the step corresponding to min(train + test) MSE value.

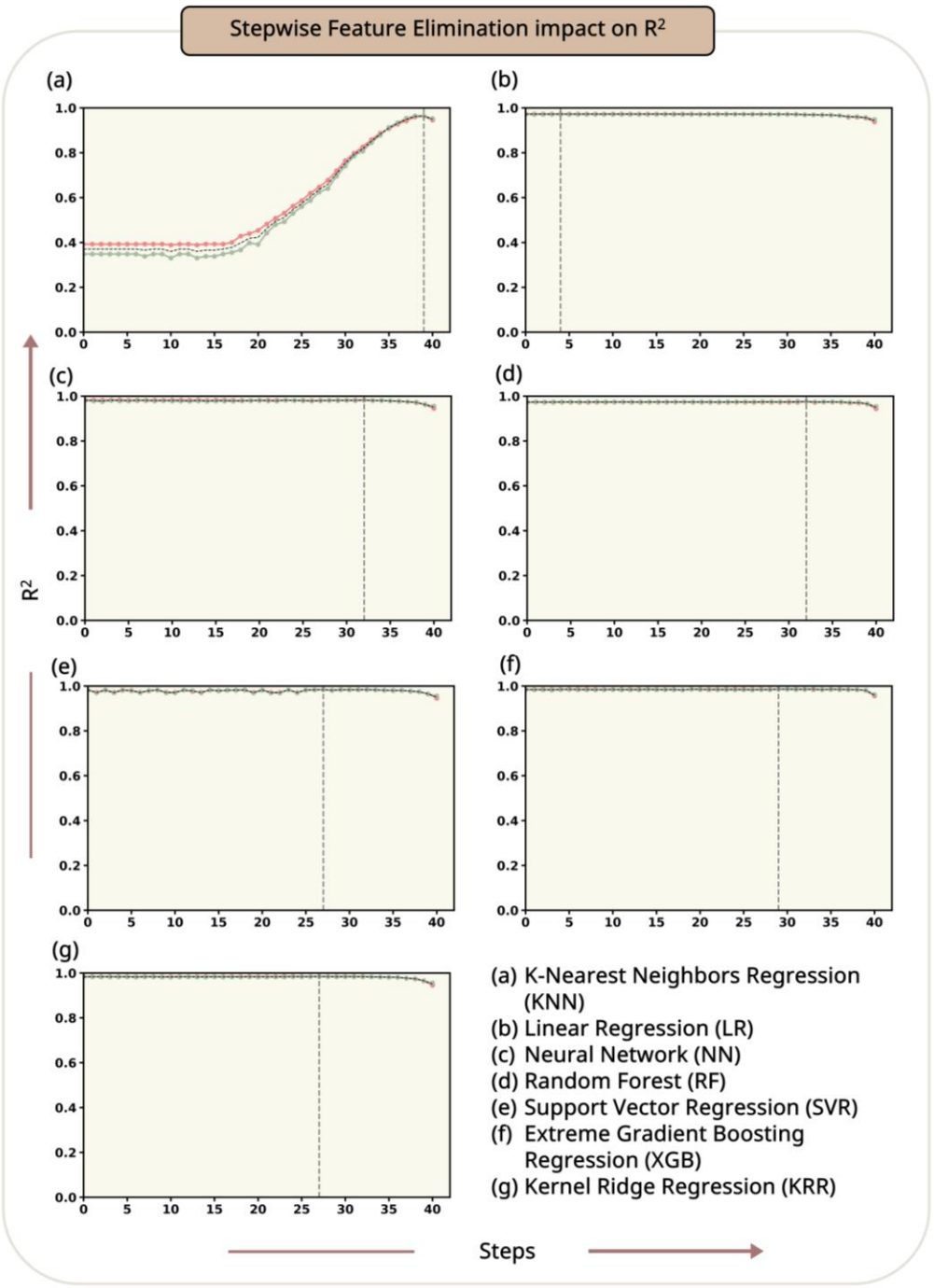

Fig S2.1. Change in coefficient of determination (R^2) values using **RDKit descriptors** for different models with reducing the number of descriptors. The red line is training R^2 and green line is testing R^2, vertical grey line is marking the step corresponding to max(train + test) R^2 value.

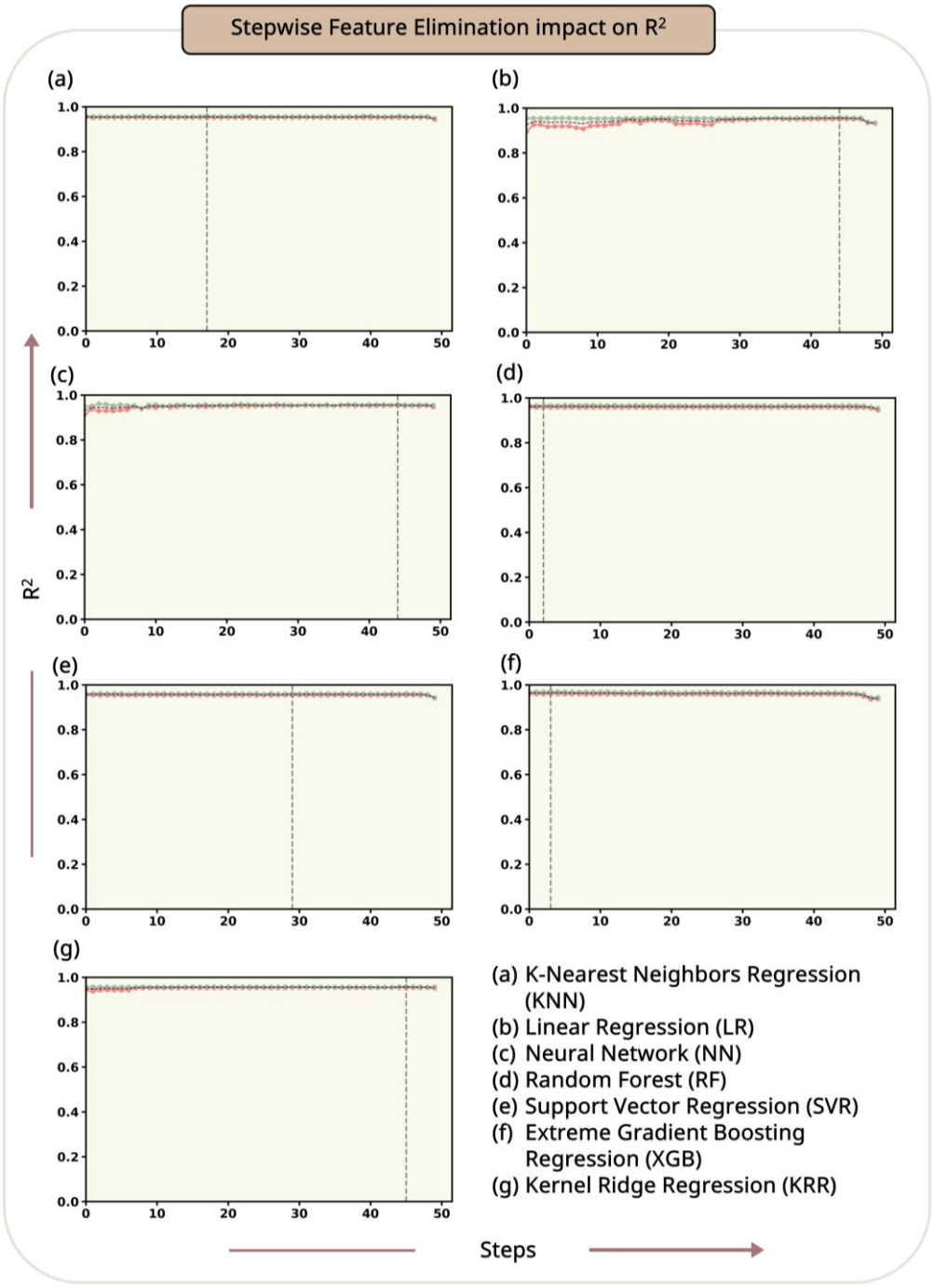

Fig S2.2. Change in coefficient of determination (R^2) values using **Sterimol descriptors** for different models with reducing the number of descriptors. The red line is training R^2 and green line is testing R^2, vertical grey line is marking the step corresponding to max(train + test) R^2 value.

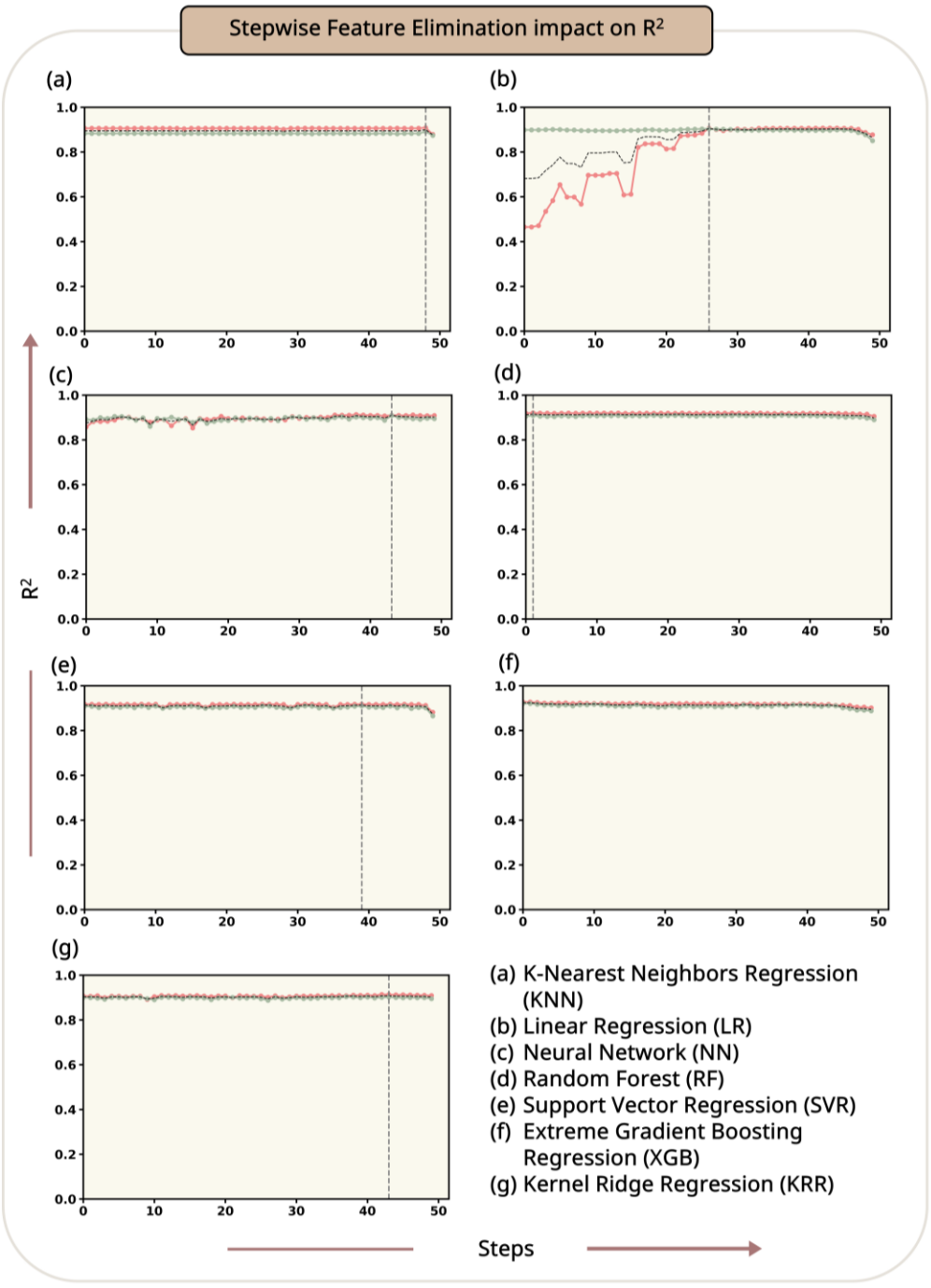

Fig S2.3. Change in coefficient of determination (R^2) values using **Sterimol descriptors without Be complexes** for different models with reducing the number of descriptors. The red line is training R^2 and green line is testing R^2, vertical grey line is marking the step corresponding to max(train + test) R^2 value.

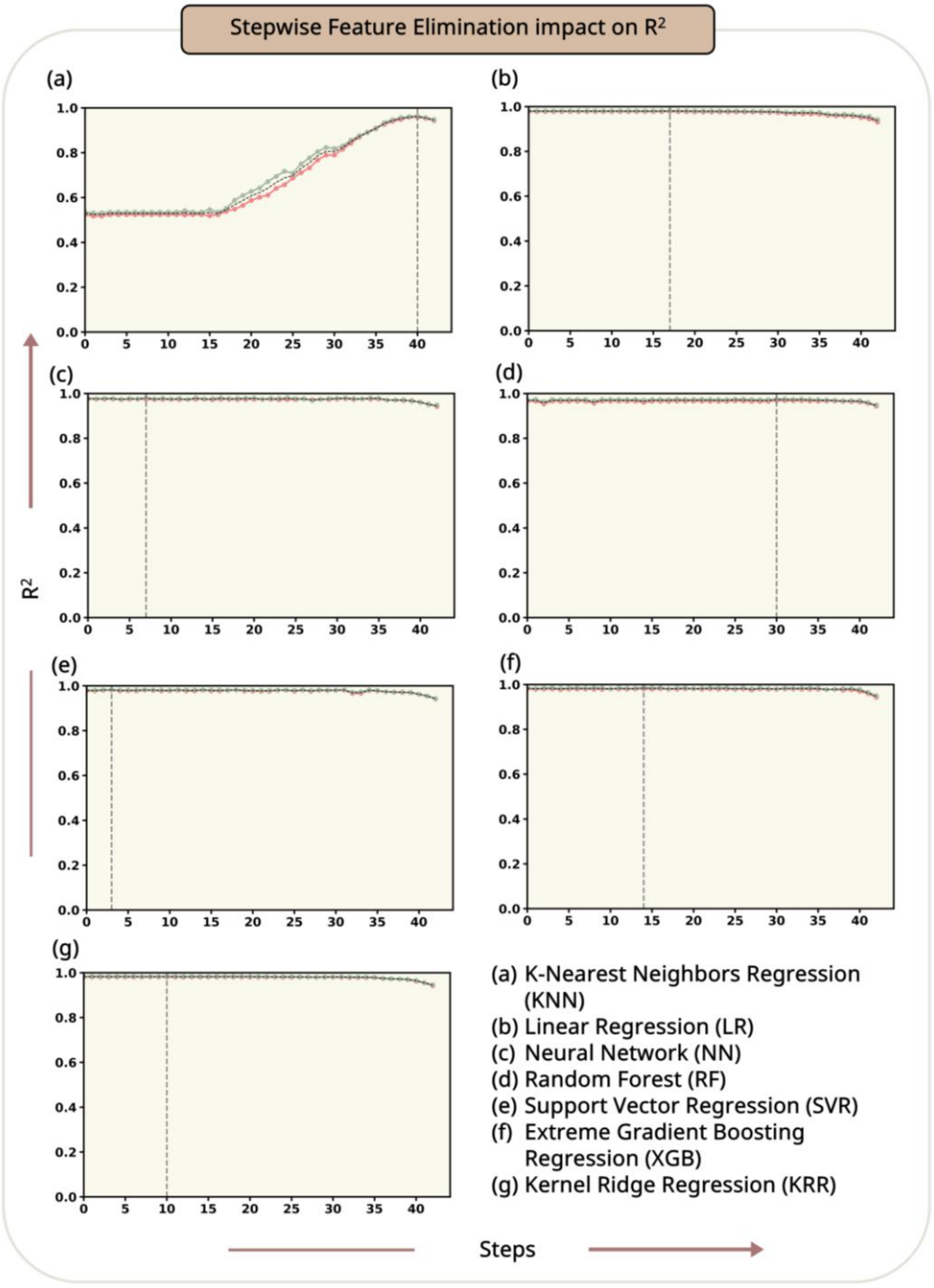

Fig S2.4. Change in coefficient of determination (R^2) values using **RDKit + Sterimol descriptors** for different models with reducing the number of descriptors. The red line is training R^2 and green line is testing R^2, vertical grey line is marking the step corresponding to max(train + test) R^2 value.

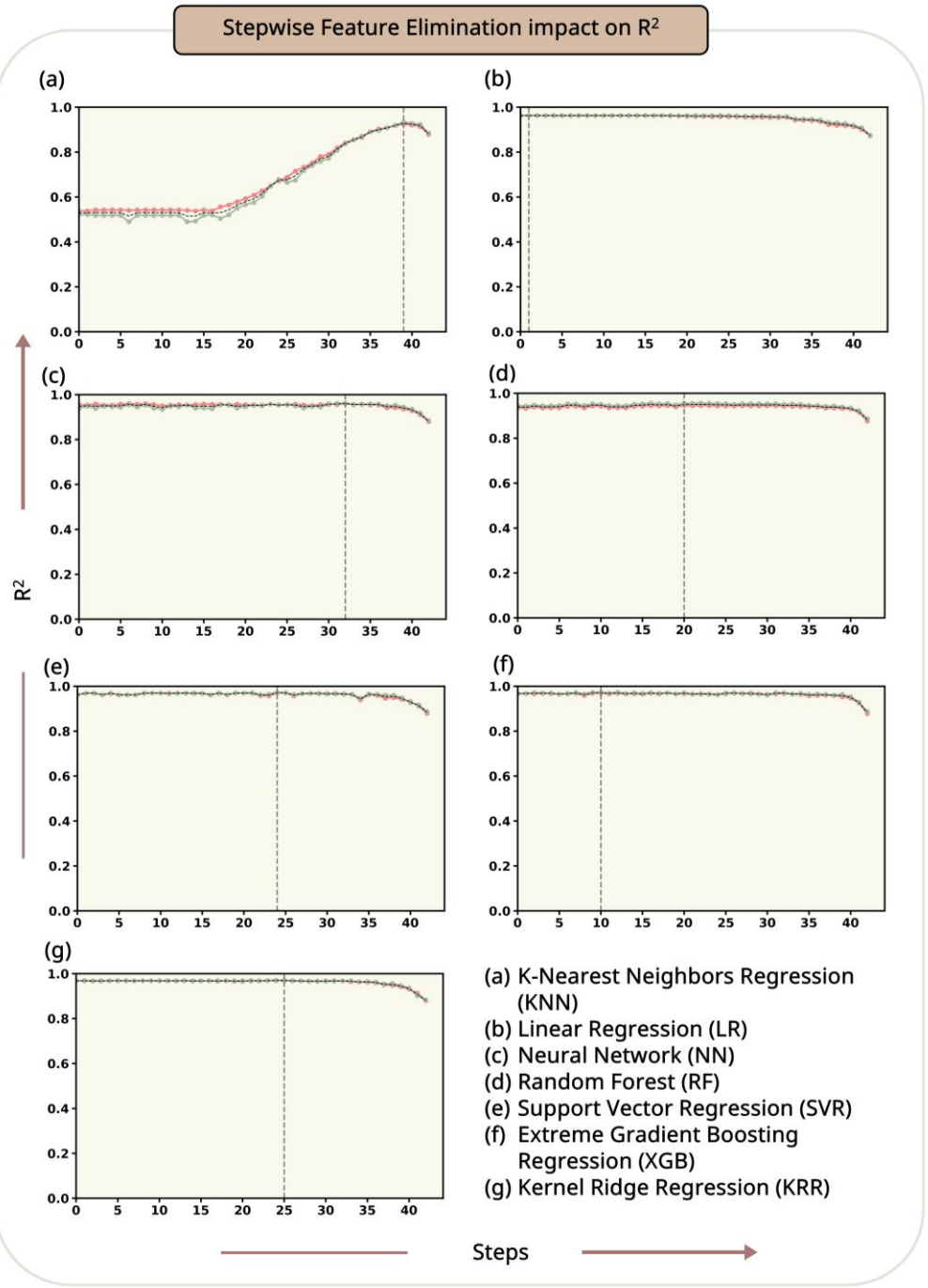

Fig S2.5. Change in coefficient of determination (R^2) values using **RDKit + Sterimol descriptors without Be complexes** for different models with reducing the number of descriptors. The red line is training R^2 and green line is testing R^2, vertical grey line is marking the step corresponding to max(train + test) R^2 value.

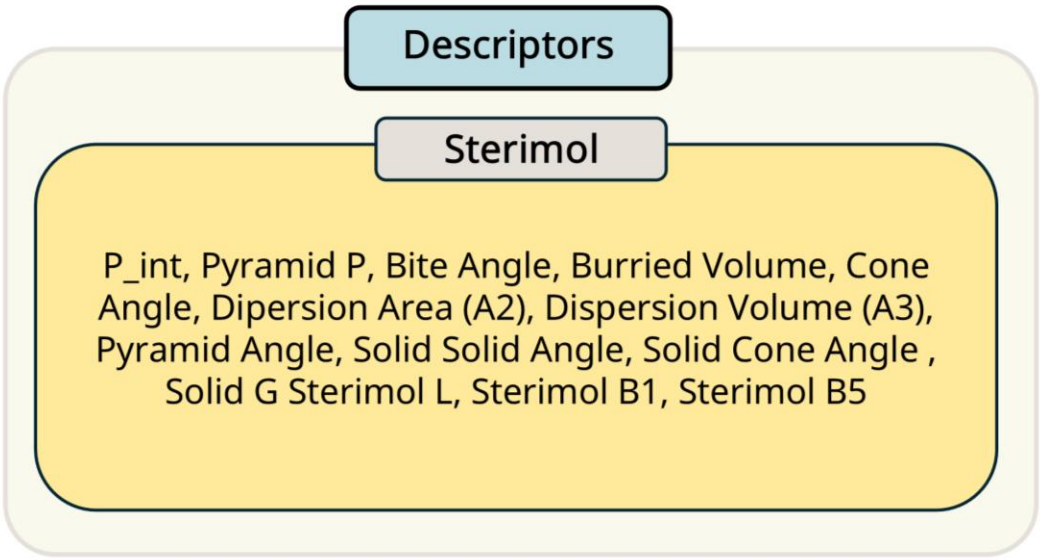

Fig S3. Sterimol used in training ML models.

```
     82 if len(within) > 0:
     83     atom_string = " ".join([str(i) for i in within])
---> 84     raise ValueError("Atoms within vdW radius of central atom:", atom_string)
     86 # Set up coordinate array and translate coordinates
     87 coordinates -= coordinates[atom_1 - 1]

ValueError: ('Atoms within vdW radius of central atom:', '12')
```

Fig S4. Error for Be complexes in Morfeus library.

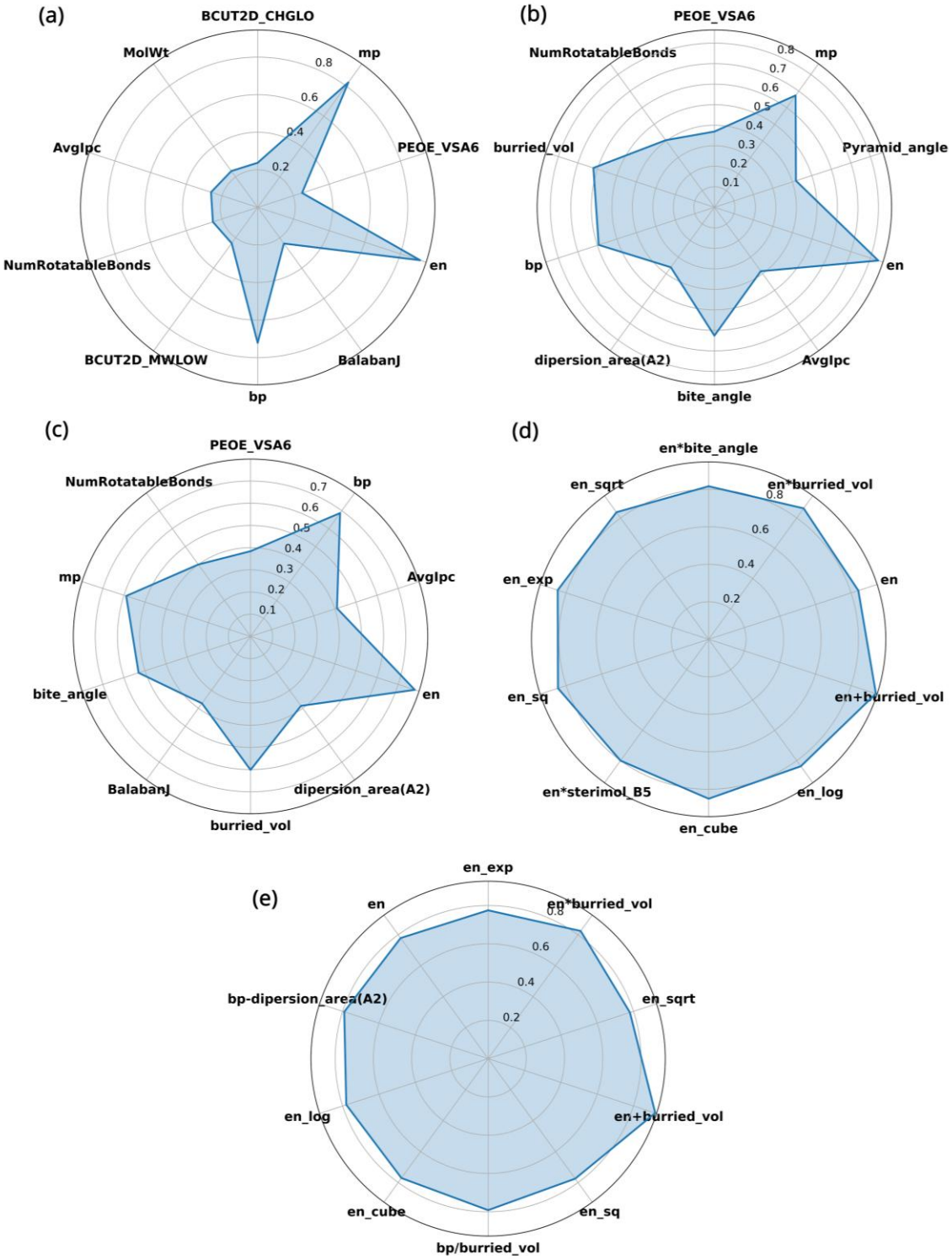

Fig S5. Sure independence screening based on distance correlation for (a) RDKit (b) RDKit + Sterimol (c) RDKit + Sterimol without Be complexes, (d) Sterimol, (e) Sterimol without Be complexes descriptor sets.

S6.1 Hyperparameter for Neural Network

```
PARAM_GRID = {
  "NN__HIDDEN_LAYER_SIZES": [(50,), (100,), (100, 50), (100, 100)],
  "NN__ACTIVATION": ["RELU", "TANH"],
  "NN__ALPHA": [1E-5, 1E-3, 1E-2],
  "NN__LEARNING_RATE_INIT": [0.001, 0.01],
}
```

S6.2 Hyperparameters for Random Forest

```
  PARAMS = {
    "N_ESTIMATORS": TRIAL.SUGGEST_INT("N_ESTIMATORS", 100, 1000),
    "MAX_DEPTH": TRIAL.SUGGEST_INT("MAX_DEPTH", 2, 30),
    "MIN_SAMPLES_SPLIT": TRIAL.SUGGEST_INT("MIN_SAMPLES_SPLIT", 2, 20),
    "MIN_SAMPLES_LEAF": TRIAL.SUGGEST_INT("MIN_SAMPLES_LEAF", 1, 10),
    "MAX_FEATURES": TRIAL.SUGGEST_CATEGORICAL("MAX_FEATURES", ["SQRT", "LOG2", NONE]),
    "BOOTSTRAP": TRIAL.SUGGEST_CATEGORICAL("BOOTSTRAP", [TRUE, FALSE]),
    "RANDOM_STATE": 42,
    "N_JOBS": -1
  }
```

S6.3 Hyperparameters for Extreme Gradient Boosting

```
  PARAMS = {
    "N_ESTIMATORS": TRIAL.SUGGEST_INT("N_ESTIMATORS", 100, 1000),
    "MAX_DEPTH": TRIAL.SUGGEST_INT("MAX_DEPTH", 2, 10),
    "LEARNING_RATE": TRIAL.SUGGEST_FLOAT("LEARNING_RATE", 0.01, 0.3, LOG=TRUE),
    "SUBSAMPLE": TRIAL.SUGGEST_FLOAT("SUBSAMPLE", 0.5, 1.0),
    "COLSAMPLE_BYTREE": TRIAL.SUGGEST_FLOAT("COLSAMPLE_BYTREE", 0.5, 1.0),
    "GAMMA": TRIAL.SUGGEST_FLOAT("GAMMA", 0.0, 5.0),
    "REG_ALPHA": TRIAL.SUGGEST_FLOAT("REG_ALPHA", 1E-8, 10.0, LOG=TRUE),
```

```
    "REG_LAMBDA": TRIAL.SUGGEST_FLOAT("REG_LAMBDA", 1E-8, 10.0, LOG=TRUE),
    "MIN_CHILD_WEIGHT": TRIAL.SUGGEST_INT("MIN_CHILD_WEIGHT", 1, 10),
    "RANDOM_STATE": 42,
    "N_JOBS": -1
  }
```

S6.4 Hyperparameters for Support Vector Regression

```
PARAMS = {"C": TRIAL.SUGGEST_FLOAT("C", 1E-2, 1E2, LOG=TRUE),
  "GAMMA": TRIAL.SUGGEST_FLOAT("GAMMA", 1E-4, 1, LOG=TRUE),
  "KERNEL": TRIAL.SUGGEST_CATEGORICAL("KERNEL", ["RBF","LINEAR"])
}
```

S6.5 Hyperparameters for K-Nearest Neighbors Regression

```
PARAMS = {
  "N_NEIGHBORS": TRIAL.SUGGEST_INT( "N_NEIGHBORS", 2, 50),
  "WEIGHTS": TRIAL.SUGGEST_CATEGORICAL( "WEIGHTS", ["UNIFORM","DISTANCE"]),
  "P": TRIAL.SUGGEST_INT("P", 1, 2),
  "LEAF_SIZE": TRIAL.SUGGEST_INT("LEAF_SIZE", 10, 60)
}
```

S6.6 Hyperparameters for Kernel Ridge Regression

```
PARAMS = { "ALPHA": TRIAL.SUGGEST_FLOAT("ALPHA", 1E-4, 10, LOG=TRUE),
"KERNEL": TRIAL.SUGGEST_CATEGORICAL("KERNEL", ["LINEAR", "RBF", "POLYNOMIAL"]),
"GAMMA": TRIAL.SUGGEST_FLOAT("GAMMA", 1E-4, 10, LOG=TRUE),
"DEGREE": TRIAL.SUGGEST_INT("DEGREE", 2, 5)
}
```

S7 – Impact of substitution effect on binding energy

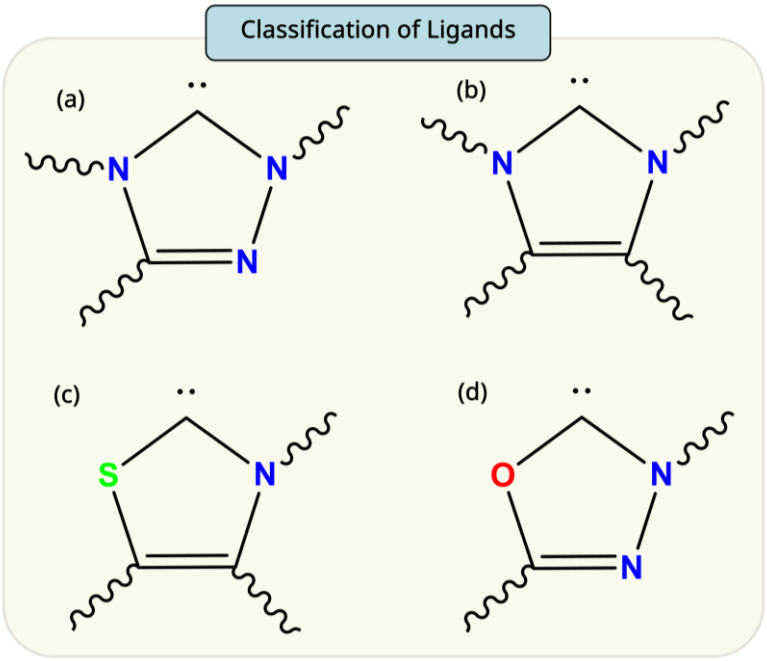

Fig S7.1. Classification of N-heterocyclic Carbenes based on ring containing carbene center.

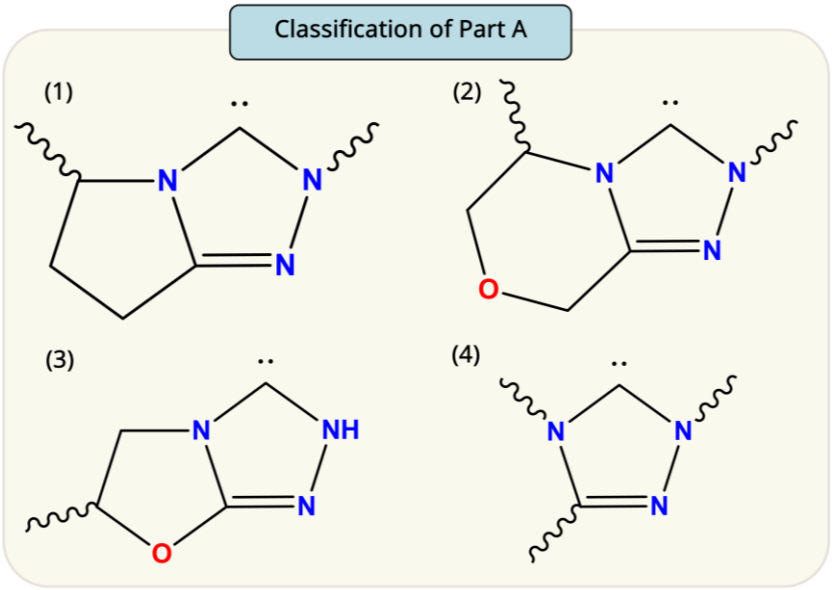

Fig S7.2. Classification of Part (a) of N-heterocyclic Carbenes based on substituted ring at ring containing carbene center.

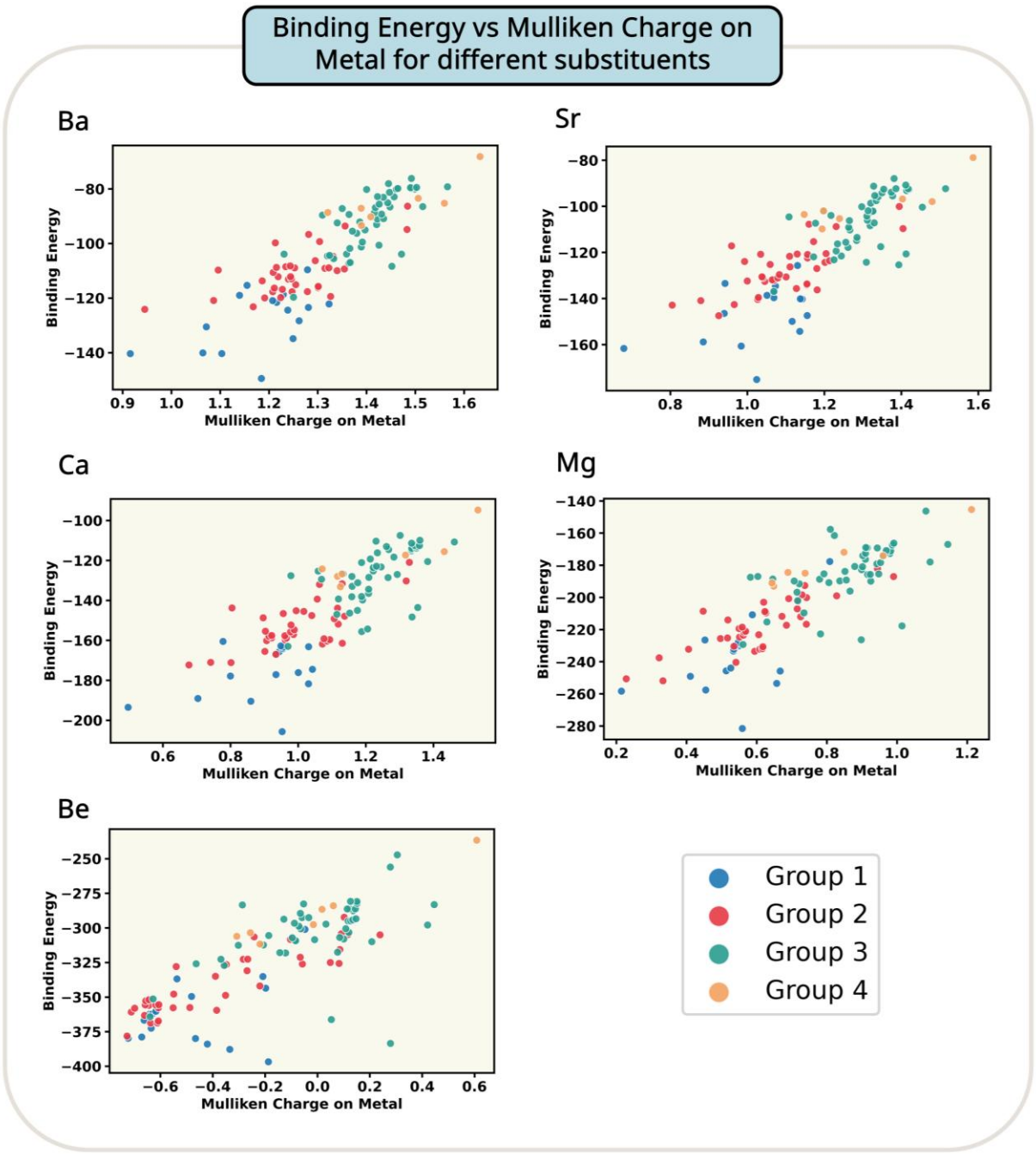

Fig S7.3. Plot of binding energy vs Mulliken charge on metal for different group of substituents for part (1) of part (a).

To examine the effect of ligand substitution, we considered only those complexes whose corresponding ligands formed complexes with all three metals, resulting in a subset of 227 ligands. These ligands were first categorized based on the heterocyclic ring containing the carbene center as depicted in the **figure S6.1**, yielding four structural classes: CNXNC (27), CNXNN (140), CNXSC (41), and COXNN (19). To obtain a clearer understanding of the substituent effects, the CNXNN class was further divided into three subgroups according to the nature of the ring fused

with the carbene containing ring, resulting in type1 (fused ring is cyclopentane), type2 (fused 6 member ring with O as heteroatom in backbone) and type3 (fused 5 member ring with O as heteroatom in backbone) and type4 (others) is remaining category as shown in the **Figure S6.2**.

We limited our study to type1 of CNXNN class, as this subgroup contains the most complexes, 107 in number. Withing this subset, we have checked substitutions in two atom positions N and C as shown in **Figure S6.2**, highlighted with pink and yellow respectively. Based on the nature of these substituents, the ligands were grouped into four categories: (i) systems containing two aromatic rings with at least one aromatic group at the second position, (ii) combinations of an aromatic group with either an electron-donating group (EDG) or electron-withdrawing group (EWG), (iii) combinations of an aromatic group with a non-aromatic (NAR) group, and (iv) other miscellaneous substituent patterns. We observed that the binding energy correlates strongly with the Mulliken charge on the metal center, and the above substituent categories exhibit a clear trend. Complexes belonging to category (i) display the most negative BE, followed by category (ii) and category), as shown in **Figure S6.3**. However, this correlation was found to be slightly weaker in the case of Be-containing complexes, likely due to the distinct electronic characteristics and smaller ionic radius of Be compared to the other metals.

S8 – Different structures of carbene used in benchmark study

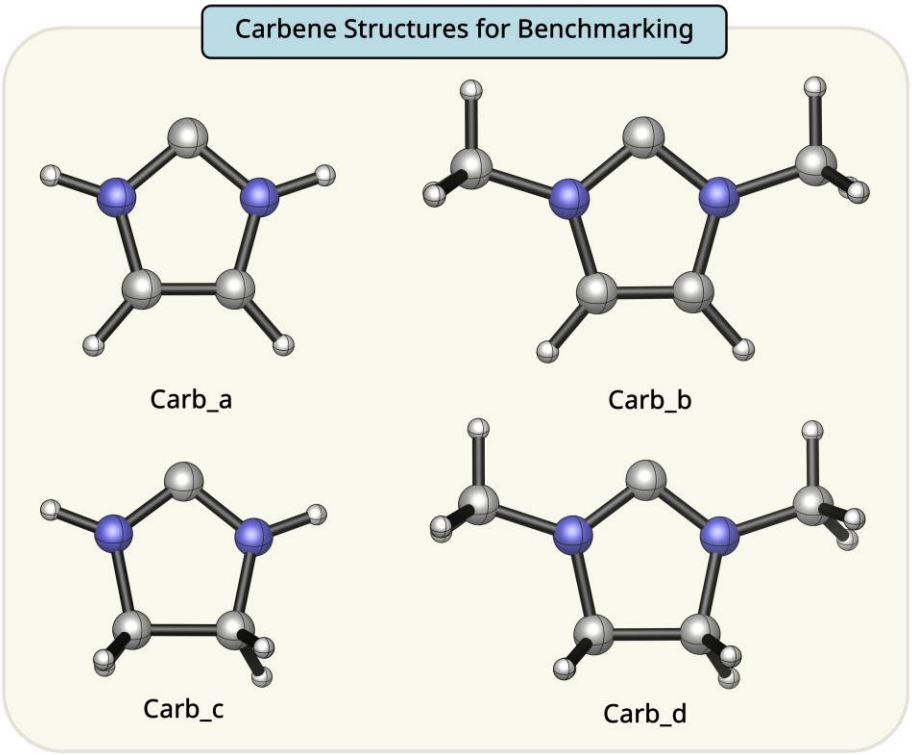

Fig S8. Different structures of carbene used in benchmark study.