**Übersicht**

Dieser Bericht soll konkret die bisherigen Ergebnisse zum Training meines ersten ML-Modells (und der zugehörigen Pipeline) zusammenfassen:

1. Die Transformation der Matrizen in ein “Universal Basis Set” (UBS) funktioniert gut, für e.g. pcseg-1 und alle Elemente von H bis Ar braucht es ca. 30 UBS-Basisfunktionen um einen relativen Rekonstruktions-Fehler in der Größenordnung 1e-14 zu erzielen.
2. Die kanonische Ausrichtung identischer Moleküle funktioniert auch ganz passabel: Berechnung der Dichte via SCF für Fe(CO)2(NO)2 in unterschiedlicher Ausrichtung und dann Transformation der Dichtematrizen in die kanonische Orientierung liefert einen absoluten Fehler der Matrixelemente in der Größenordnung 1e-4.
3. Modelle, die basierend auf der lokalen Umgebung (“Sub-Molekül”, modelliert als Graph) einzelne Blöcke der Fock-Matrix vorhersagen funktionieren erstaunlich gut.
4. Die besten auf diese Weise generierten Density-Guesses erreichen etwa die selbe Performance (in Iterationen bis zur Konvergenz) wie der beste PySCF-Guess (minao), im Schnitt ist die Performance jedoch immer noch schlechter.

Meine Graph-NN Architektur zur Vorhersage von einzelnen Fock-Matrix-Blöcken scheint prinzipiell gut zu funktionieren, deshalb würde ich vorerst dabei bleiben. Im letzten Schritt (bei der Rücktransformation von kanonischer in nicht-kanonische Orientierung der Molekülgeometrie, alles ausgedrückt im originalen Basissatz) scheint sich der relative Fehler zwischen Vorhergesagter und erwarteter Fock-Matrix plötzlich zu verzehnfachen. Da ich die Transformation an sich gut getestet habe gehe ich davon aus, dass sich irgendwo anders noch ein Bug eingeschlichen hat. Prinzipiell sollte das Problem aber lösbar sein. Da meine besten Guesses bereits jetzt auf minao-Niveau liegen ist davon auszugehen, dass meine Guesses nach Behebung des Problems im Schnitt besser sein werden.

**Pipeline**

- Gearbeitet wurde mit den C7O2H10-Isomeren des QM9-Datensatzes
- Jede Overlap/Density/Fock-Matrix wurde zuerst blockweise (basierende auf der lokalen Umgebung der Atome i und j) in eine kanonische Ausrichtung transformiert
- Die transformierten Matrizen wurden dann blockweise im Universal Basis Set ausgedrückt (aus rechteckigen Blöcken werden quadratische mit einheitlicher Größe)
- Daraus wurde ein Trainings-Datensatz zur Vorhersage von einzelnen Fock-Matrix-Blöcken erstellt
- Das Training des Modells erfolgt in dieser speziellen Repräsentation der Daten. Es werden zwei Modelle trainiert: jeweils für diagonal-Blöcke und off-diagonal-Blöcke. Ein einzelnes Modell kann flexibel unabhängig von den involvierten Atom-Spezies und der Anzahl der benachbarten Atome einen Block vorhersagen. Als beste Loss-Function hat sich bisher der RMSE erwiesen.
- Die gesamte Guess-Matrix wird dann aus individuellen Blöcken assembliert und danach blockweise zurück-transformiert.

**Model Input**

- Die nähere Umgebung der beiden Atome i und j (für die ein Block vorhergesagt werden soll) wird als Graph der benachbarten Atome dargestellt
- Welche Atome als Nachbarn infrage kommen wird über den räumlichen Überlapp der Basisfunktionen bestimmt: wird der Überlapp zwischen Basisfunktionen zweier Atome zu gering, so werden diese nicht mehr als Nachbarn gezählt
- Die Node-Features beinhalten Kennwerte wie den kovalenten Radius, die Ladung, Anzahl der Valenzelektronen etc.
- Die Edge-Features bestehen aus den zugehörigen Overlap-Blöcken zwischen den beteiligten Atomen

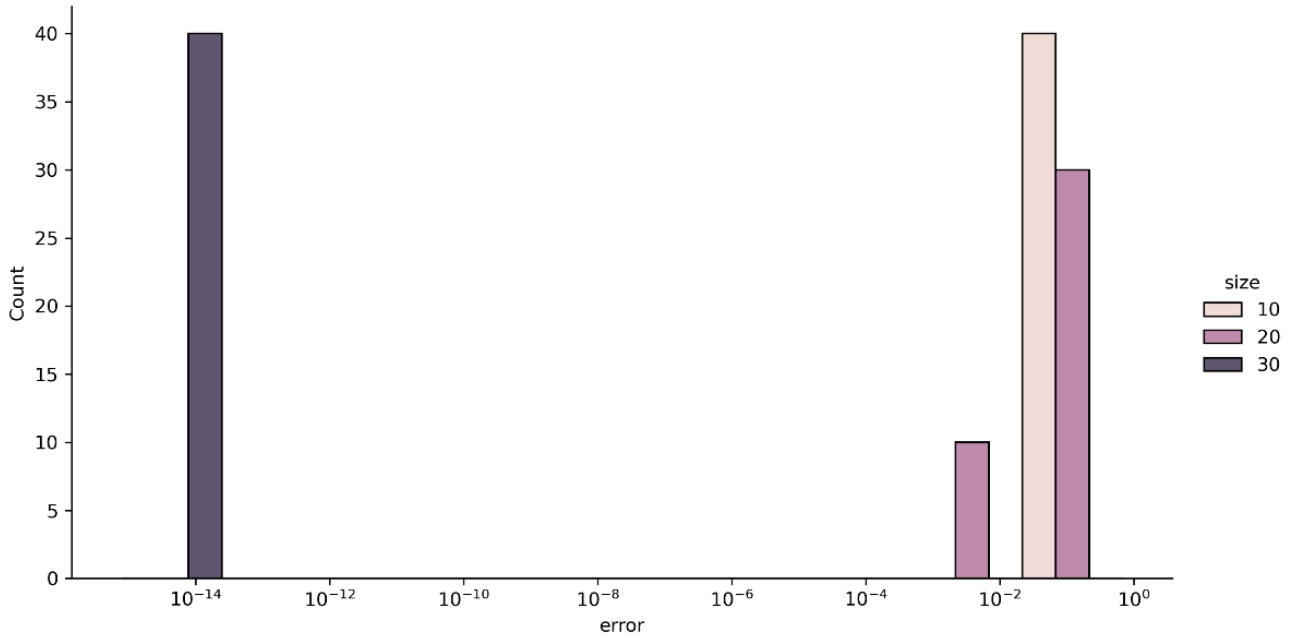


Abbildung 1: Histogramm der relativen Rekonstruktions-Fehler der Overlap/Density/Fock-Matritzen für einige Moleküle aus dem Datensatz, abhängig von der Größe des Universal Basis Set. Hier wurde der 6-31G(2df,p) Basissatz (nur für H, C und O) verwendet - aber auch bei höherer Anzahl der der originalen Basisfunktionen (wie etwa pcseg-1) liegt die Anzahl der benötigten UBS-Basisfunktionen im Bereich von 30-50.

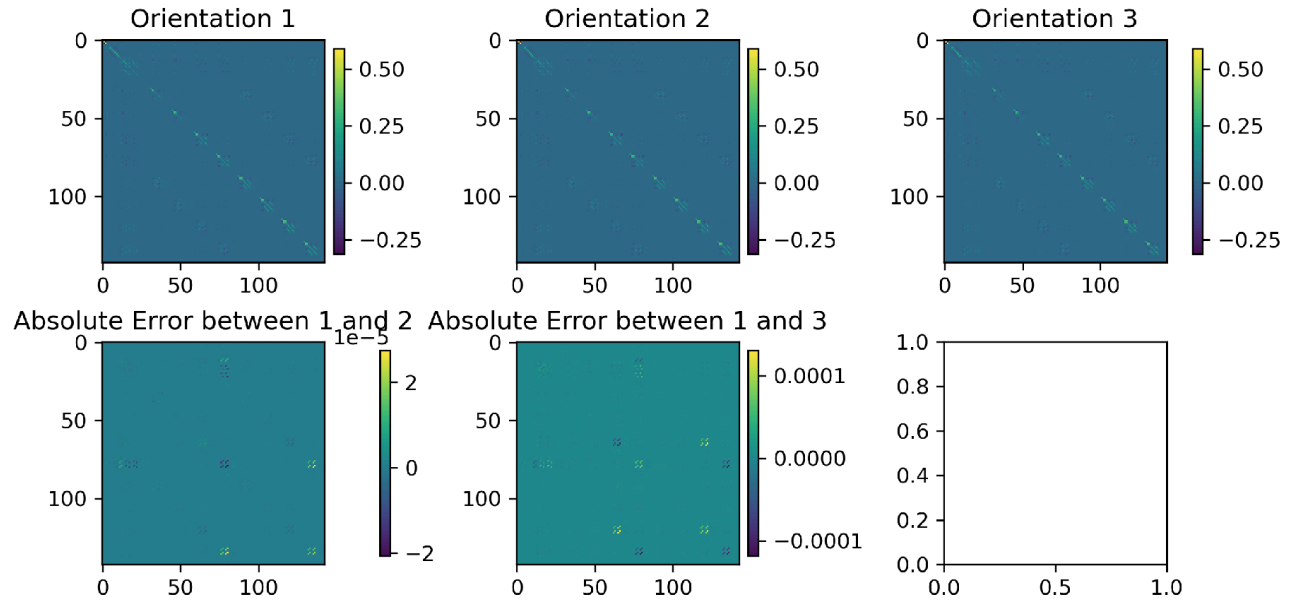


Abbildung 2: Kanonische Repräsentation der Dichtematritzen von Fe(CO)2(NO)2, welche ursprünglich in unterschiedlicher räumlicher Ausrichtung berechnet wurden. Der Absolute Fehler zwischen den transformierten Matritzen liegt bei etwa 1e-4.

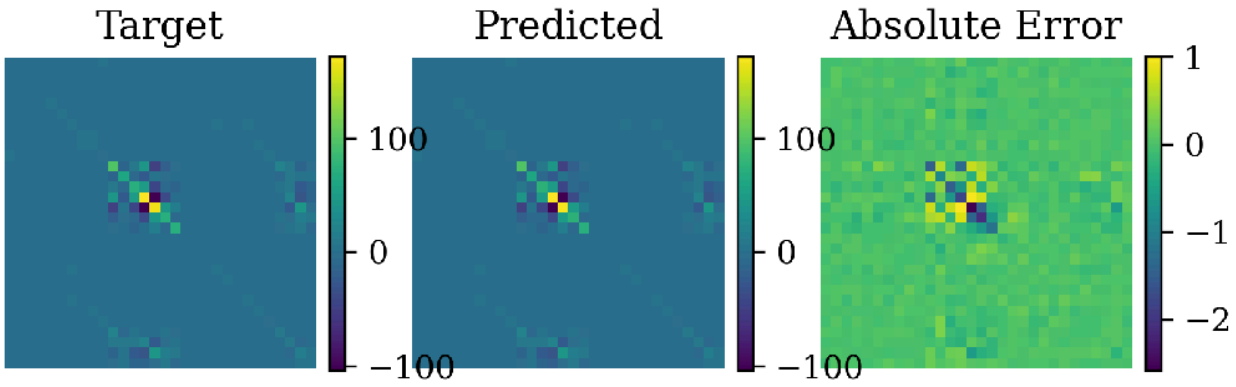


Abbildung 3: Schlechtester Guess des diagonal-block-Predictors für ca. 2000 Blöcke aus dem Validierungsdatensatz (100 nicht im Trainingsdatensatz enthaltene Moleküle). Die hier gezeigten Blöcke sind in kanonischer Orientierung und ausgedrückt in der Universal Basis.

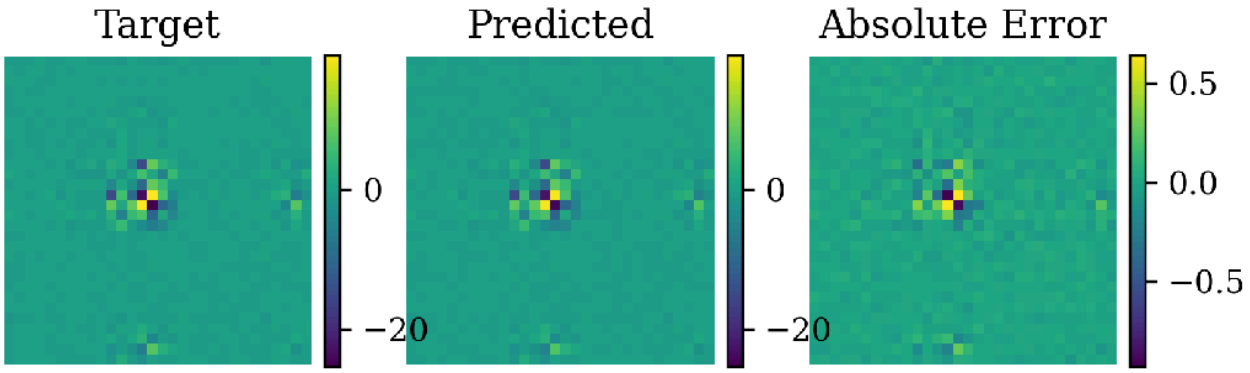


Abbildung 4: Schlechtester Guess des off-diagonal-block-Predictors für ca. 8000 Blöcke aus dem Validierungsdatensatz (100 nicht im Trainingsdatensatz enthaltene Moleküle). Die hier gezeigten Blöcke sind in kanonischer Orientierung und ausgedrückt in der Universal Basis.

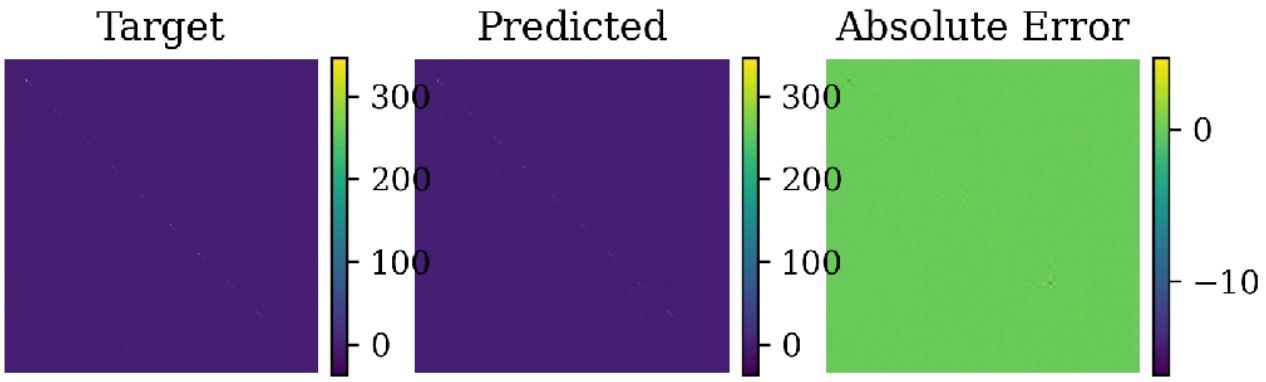


Abbildung 4: Schlechtester Gesamt-Guess der Fock-Matrix aus dem Validierungsdatensatz (100 nicht im Trainingsdatensatz enthaltene Moleküle). Die hier gezeigten Blöcke sind in kanonischer Orientierung und im originalen Basissatz ausgedrückt. Der Absolute Fehler ist vergleichsweise gering.

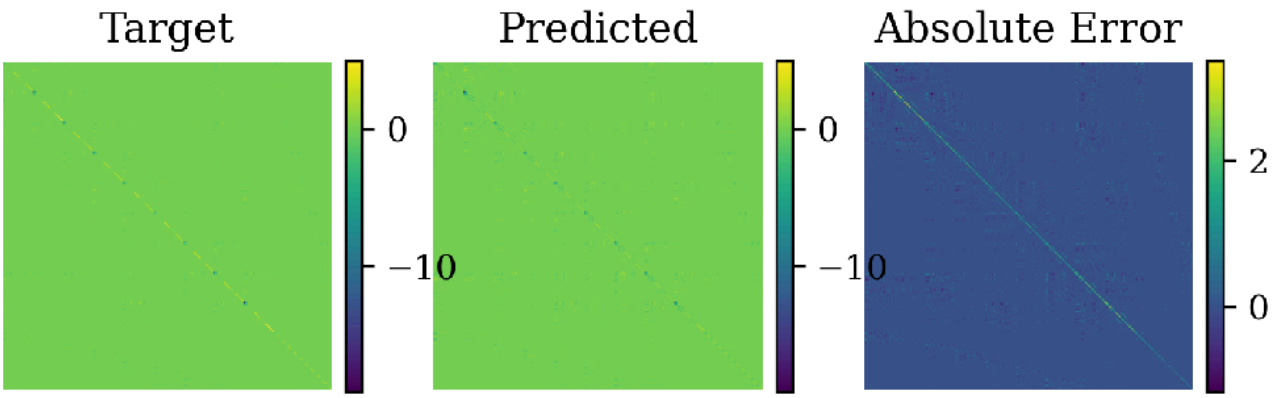


Abbildung 4: Schlechtester Finaler Guess der Fock-Matrix aus dem Validierungsdatensatz (100 nicht im Trainingsdatensatz enthaltene Moleküle). Die hier gezeigten Blöcke sind in nicht-kanonischer Orientierung und im originalen Basissatz ausgedrückt. Der Absolute Fehler ist relativ zum erwarteten Wertebereich jetzt plötzlich um den Faktor 10 höher als in Abbildung 4.

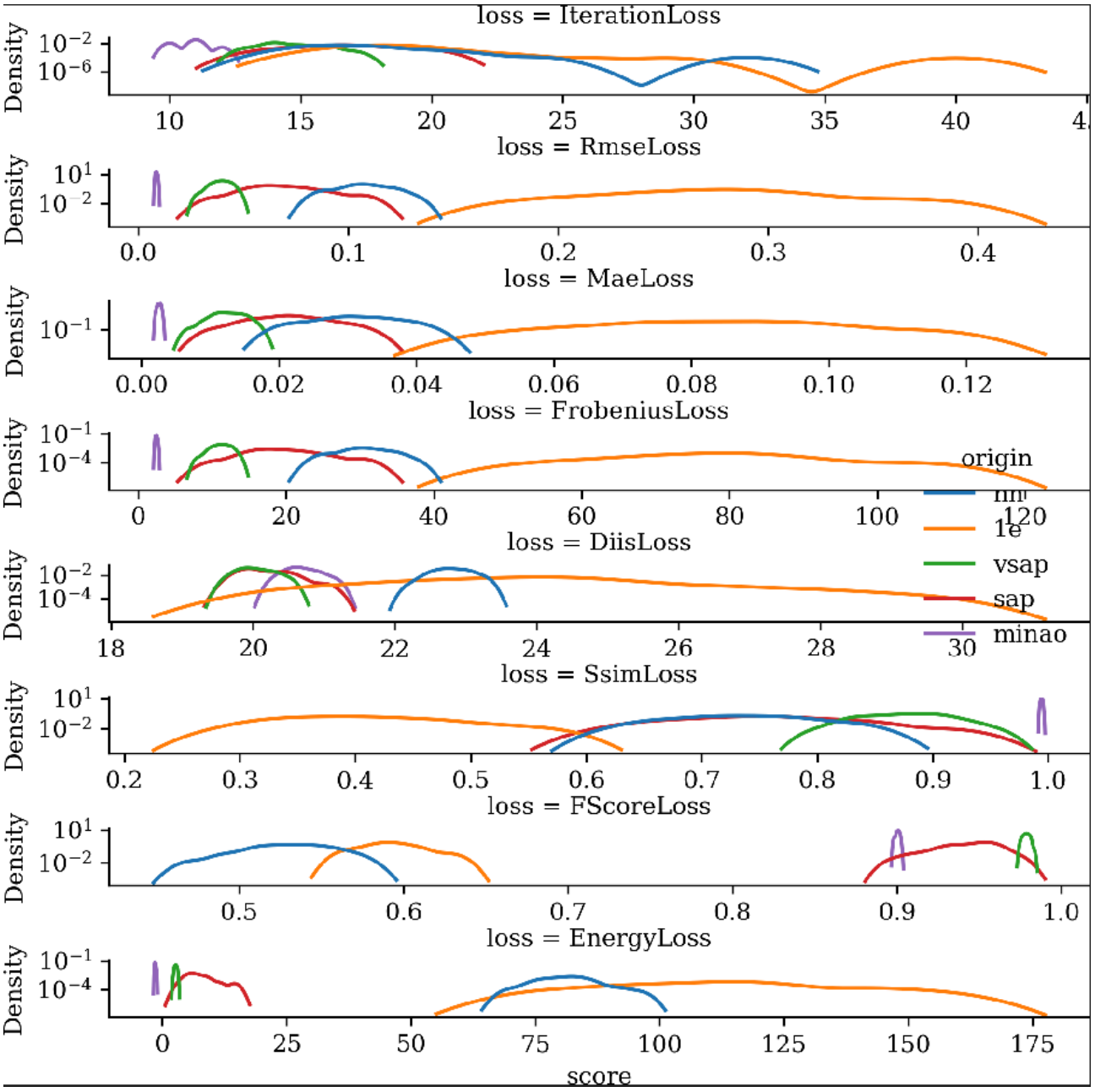


Abbildung 5: Vergleich der finalen Guesses für 100 Moleküle aus dem Validierungs-Datensatz. “nn” (blau) ist der ML-basierte Guess, “minao” (violett) der beste PySCF-Guess. Die hier gezeigten Kennzahlen sind (abgesehen von “IterationLoss”) nicht notwendigerweise gute Indikatoren für die Performance, und den Absolutwerten für beispielsweise DiisLoss und EnergyLoss kann bei der derzeitigen Implementierung keine gängige physikalische Bedeutung beigemessen werden!