# 1 Introduction

As part of James E. Avery's efforts to develop an efficient screening pipeline for fullerenes and potentially fulleroids we will in this report detail our efforts in porting parts of the xtb program by Grimme et al to SYCL code. The goal is a highly optimised and fully lockstep-parallel implementation of the electronic structure calculations from the GFN2-xTB method. A previous thesis by la Cour provides an efficient lockstep-parallel implementation of a forcefield method for computing geometric structures of fullerenes, which we will take to be our input.

The mentioned screening pipeline would enable the search of entire isomer spaces for fullerenes with certain properties such as a low lowest energy state indicating a stable isomer.

A fullerene is a molecule consisting only of carbon atoms connected in 12 pentagons and enough hexagons to create a hollow structure. As we increase the number of atoms the isomer space quickly grows leading to very slow search times. We aim to provide a quick and relatively accurate method for discarding large unpromising parts of the isomer space before searching with more accurate methods.

Fulleroids are essentially an extension to fullerenes where we allow n-gons instead of only penta- and hexagons, as long as we can still create a closed shape.

After a literature review we settled on the Geometry, Frequency, Noncovalent, extended Tight Binding (GFNn-xTB) family of methods as they are relatively accurate and quite fast at predicting electronic structures to a reasonable accuracy. We hope that this inherent speed will aid in getting good though-put after the transformation to a lockstep-parallel version. The GFNn-xTB metods come in iterations. GFN1 is the first and lays the ground work for the later iterations. It does however rely on element pairwise specific constants. In GFN2 this has been changed in favour of only element specific parameters. GFN0 is a more approximate and faster version of GFN2. And GFN-FF takes this trade-off further as this is a forcefield method which is parametrised using the insights (and parameters) gained from the other GFN iterations.

Forcefield methods save on computing all the pairwise interactions between atoms in a molecule and instead use efficient rules to lump atoms together in predictable clumps which then interact with other clumps. This can save tremendous effort.

Specifically GFN2 seemed most promising for our purposes as it is more accurate than GFN0 and simpler than GFN1, and if it is not fast enough would be relatively easy to then implement GFN0. GFN-FF was not considered suitable due to us wanting to see if it could be fast enough without defaulting to a forcefield method.

Lockstep-parallelisation is a paradigm best suited for GP-GPU. It takes advantage of the fact that GPUs operate more efficiently when all the cores are doing the same operations in a predictable fashion. This essentially is a step beyond data parallelism. We are not only operating on the same data across

cores, but also doing the exact same steps. This means no conditionals with a data dependent evaluation. It is fine to have a loop that runs five times, opposed to say `data[coreId]` times.

# 2 Extended Hückel Theory Matrix for GFN2-xTB

$$\begin{aligned} H_{\mu\nu}^{EHT} = & \frac{1}{2} K_{AB}^{ll'} S_{\mu\nu} (H_{\mu\mu} + H_{\nu\nu}) \\ & \cdot X(EN_A, EN_B) \\ & \cdot \Pi(R_{AB}, l, l') \\ & \cdot Y(\zeta_l^A, \zeta_{l'}^B), \forall \mu \in l(A), \nu \in l'(B) \end{aligned} \tag{1}$$

where $\mu$ and $\nu$ are AO indecies, $l$ and $l'$ index shells. Both AO's are associated with an atom labled A and B. $K_{AB}^{ll'}$ is a element and shell specific fitted constant however, in GFN2 it only depends on the shells. $S_{\mu\nu} = \langle \phi_\mu | \phi_\nu \rangle$ is just the overlap of the orbitals. In GFN2 $H_{\kappa\kappa} = h_A^l - \delta h_{CN_A'}^l CN_A'$ where $CN_A'$ is the modified GFN2-type Coordinate Number for the element of atom A.

$$\begin{aligned} CN_A^{'} = & \sum_{B \neq A}^{N_{\text{atoms}}} (1 + e^{-10(4(R_{A,\text{cov}} + R_{B,\text{cov}})/3R_{AB} - 1)})^{-1} \\ & \times (1 + e^{-20(4(R_{A,\text{cov}} + R_{B,\text{cov}} + 2)/3R_{AB} - 1)})^{-1} \end{aligned} \tag{2}$$

$h_A^l$ and $\delta h_{CN_A'}^l$ are both fitted constants. $EN_A$ is the electronegativity of the element of atom A, given in the original `xtb` code.

$$X(EN_A, EN_B) = 1 + k_{EN} \Delta EN_{AB}^2 \tag{3}$$

$$k_{EN} = 0.02 \text{ in GFN2} \tag{4}$$

$$\Delta EN_{AB}^2 = (EN_A - EN_B)^2 \tag{5}$$

$$\Pi(R_{AB}, l, l') = \left(1 + k_{A,l}^{\text{poly}} \left(\frac{R_{AB}}{R_{\text{cov},AB}}\right)^{\frac{1}{2}}\right) \left(1 + k_{B,l'}^{\text{poly}} \left(\frac{R_{AB}}{R_{\text{cov},AB}}\right)^{\frac{1}{2}}\right) \tag{6}$$

$R_{\text{cov},AB}$ are the summed covalent radii ($R_{\text{cov},A} + R_{\text{cov},B}$), e.g. $R_{\text{cov},H} = 0.32$, $R_{\text{cov},C} = 0.75$ are given in the original `xtb` code. $k_{A,l}^{\text{poly}}$ and $k_{B,l'}^{\text{poly}}$ are element and shell specific constants.

$$Y(\zeta_l^A, \zeta_{l'}^B) = \left(\frac{2\sqrt{\zeta_l^A \zeta_{l'}^B}}{\zeta_l^A + \zeta_{l'}^B}\right)^{\frac{1}{2}} \tag{7}$$

Here, $\zeta_l^A$ are the STO exponents of the GFN2-xTB AO basis.
Slater Type Orbitals are defined as such:

$$\chi_{\zeta,n,l,m}(r,\theta,\varphi) = NY_{l,m}(\theta,\varphi)r^{n-1}e^{-\zeta r} \tag{8}$$

N is a normalisation constant, Y are spherical harmonic funtions, n, l, m are the quantum numbers for the AO. $r, \theta, \varphi$ are polar 3D coordinates. $\zeta$ determines the radial extent of the STO, a large value gives rise to a function that is "tight" around the nucleus and a small value gives a more "diffuse" function. This $\zeta$ is the one mentioned in the Y term of $E_{EHT}$ and is a value fitted when constructing the basis set, thus it is given to us.

# 3 Fock Matrix for GFN2-xTB

$$F_{\mu\nu}^{GFN2-xTB} = H_{\mu\nu}^{EHT} + F_{\mu\nu}^{IES+IXC} + F_{\mu\nu}^{AES} + F_{\mu\nu}^{AXC} + F_{\mu\nu}^{D4}, \quad \forall \mu \in A, \nu \in B \tag{9}$$

## 3.1 Isotropic Electrostatic and Exchange-correlation contribution

$$\begin{aligned} F_{\mu\nu}^{IES+IXC} = &-\frac{1}{2}S_{\mu\nu}\sum_C\sum_{l''}(\gamma_{AC,ll''} + \gamma_{BC,l'l''})q_{C,l''} \\ &-\frac{1}{2}S_{\mu\nu}(q_{A,l}^2\Gamma_{A,l} + q_{B,l'}^2\Gamma_{B,l'}) \end{aligned} \tag{10}$$

$l, l', l''$ being the angular momenta of the orbitals $\mu, \nu$ and each of C's orbitals.

$$\Gamma_{A,l} = K_l^\Gamma \Gamma_A \tag{11}$$

$K_l^\Gamma$ is a shell specific constant common for all elements and $\Gamma_A$ is an element specific constant.

$$\gamma_{AB,ll'} = \frac{1}{\sqrt{R_{AB}^2 + \eta_{AB,ll'}^{-2}}} \tag{12}$$

$$\eta_{AB,ll'} = \frac{1}{2}\left[\eta_A(1+k_A^l) + \eta_B(1+k_B^{l'})\right] \tag{13}$$

$q_l$ is a partial Mulliken charge. $\eta_A$ and $\eta_B$ are element-specific fit parameters, while $k_A^l$ and $k_B^{l'}$ are element-specific scaling factors for the individual shells ($k_A^l = 0$ when $l = 0$).

$$GAP_A = \sum_{l\in A} q_{A,l} \tag{14}$$

$$q_{A,l} = \sum_{l'\in B} P_{ll'}S_{ll'} = GOP_l \tag{15}$$

### 3.2 Anisotropic Electrostatic and Exchange-correlation contribution

$$\begin{aligned} F_{\mu\nu}^{AES} + F_{\mu\nu}^{AXC} &= \frac{1}{2} S_{\mu\nu} \left[ V_S(\boldsymbol{R}_B) + V_S(\boldsymbol{R}_C) \right] \\ &+ \frac{1}{2} \boldsymbol{D}_{\mu\nu}^T \left[ \boldsymbol{V}_D(\boldsymbol{R}_B) + \boldsymbol{V}_D(\boldsymbol{R}_C) \right] \\ &+ \frac{1}{2} \sum_{\alpha,\beta \in \{x,y,z\}} Q_{\mu\nu}^{\alpha\beta} \left[ V_Q^{\alpha\beta}(\boldsymbol{R}_B) + V_Q^{\alpha\beta}(\boldsymbol{R}_C) \right] \end{aligned} \tag{16}$$

$$\boldsymbol{D}_{\mu\nu}^T = \begin{pmatrix} D_{\mu\nu}^x & D_{\mu\nu}^y & D_{\mu\nu}^z \end{pmatrix} \tag{17}$$

$$\tag{18}$$

$$\begin{aligned} V_S(\boldsymbol{R}_C) = \sum_A \Bigg\{ & \boldsymbol{R}_C^T \Big[ f_5(R_{AC}) \boldsymbol{\mu}_A R_{AC}^2 - \boldsymbol{R}_{AC} 3 f_5(R_{AC}) (\boldsymbol{\mu}_A^T \boldsymbol{R}_{AC}^2) \\ & - f_3(R_{AC}) q_A \boldsymbol{R}_{AC} \Big] - f_5(R_{AC}) \boldsymbol{R}_{AC}^T \boldsymbol{\Theta}_A \boldsymbol{R}_{AC} - f_3(R_{AC}) \boldsymbol{\mu}_A^T \boldsymbol{R}_{AC} \\ & + q_A f_5(R_{AC}) \frac{1}{2} \boldsymbol{R}_C^2 \boldsymbol{R}_{AC}^2 - \frac{3}{2} q_A f_5(R_{AC}) \sum_{\alpha\beta} \alpha_{AB} \beta_{AB} \alpha_C \beta_C \Bigg\} \\ & + 2 f_{XC}^{\mu_C} \boldsymbol{R}_C^T \boldsymbol{\mu}_C - f_{XC}^{\Theta_C} \boldsymbol{R}_C^T \Big[ 3 \boldsymbol{\Theta}_C - \mathrm{Tr}(\boldsymbol{\Theta}_C) \boldsymbol{I} \Big] \boldsymbol{R}_C \end{aligned} \tag{19}$$

**QUESTION: Should this not be $\boldsymbol{R}_C^{2,T}$, in line 3, term 1?**

$$\begin{aligned} V_D(\boldsymbol{R}_C) = \sum_A \Bigg[ & \boldsymbol{R}_{AC} 3 f_5(R_{AC}) (\boldsymbol{\mu}_A^T \boldsymbol{R}_{AC}) - f_5(R_{AC}) \boldsymbol{\mu}_A R_{AC}^2 + f_3(R_{AC}) q_A \boldsymbol{R}_{AC} \\ & - q_A f_5(R_{AC}) \boldsymbol{R}_C R_{AC}^2 + 3 q_A f_5(R_{AC}) \boldsymbol{R}_{AC} \sum_\alpha \alpha_C \alpha_{AC} \Bigg] \\ & - 2 f_{XC}^{\mu_C} \boldsymbol{\mu}_C - 2 f_{XC}^{\Theta_C} \Big[ 3 \boldsymbol{\Theta}_C - \mathrm{Tr}(\boldsymbol{\Theta}_C) \boldsymbol{I} \Big] \boldsymbol{R}_C \end{aligned} \tag{20}$$

$$\begin{aligned} V_Q^{\alpha\beta}(\boldsymbol{R}_C) = & - \sum_A q_A f_5(R_{AC}) \left[ \frac{3}{2} \alpha_{AC} \beta_{AC} - \frac{1}{2} R_{AB}^2 \right] \\ & - f_{XC}^{\Theta_C} \left[ 3 \boldsymbol{\Theta}_C^{\alpha\beta} - \delta_{\alpha\beta} \sum_\alpha \boldsymbol{\Theta}_C^{\alpha\alpha} \right] \end{aligned} \tag{21}$$

$\boldsymbol{\mu}_A$ is the cumulative atomic dipole moment of atom A and $\boldsymbol{\Theta}_A$ is the corresponding traceless quadrupole moment. Traceless simply means that the sum of

the diagonal elements is 0. The curly braces and brackets are used in the same way as normal parenthesis for showing order of operations. $q_A$ is the atomic charge of atom A.

$$\Theta_A^{\alpha\beta} = \frac{3}{2}\theta_A^{\alpha\beta} - \frac{\delta_{\alpha\beta}}{2}\left(\theta_A^{xx} + \theta_A^{yy} + \theta_A^{zz}\right) \tag{22}$$

$$\theta_A^{\alpha\beta} = \sum_{l' \in A}\sum_{l} P_l \left(\alpha_A D_{ll'}^{\beta} + \beta_A D_{ll'}^{\alpha} - \alpha_A \beta_A S_{ll'} - Q_{ll'}^{\alpha\beta}\right) \tag{23}$$

$$q_A = Z_A - GAP_A \tag{24}$$

$$\mu_A^{\alpha} = \sum_{l' \in A}\sum_{l} P_{l'l}\left(\alpha_A S_{l'l} - D_{l'l}^{\alpha}\right) \tag{25}$$

$$D_{ll'}^{\alpha} = \langle\phi_l|\alpha_i|\phi_{l'}\rangle = \langle\phi_l(\alpha_i)|\alpha_i|\phi_{l'}(\alpha_i)\rangle = \int \alpha_i \phi_l^*(\alpha_i)\phi_{l'}(\alpha_i)d\alpha_i \tag{26}$$

$$Q_{ll'}^{\alpha\beta} = \langle\phi_l|\alpha_i\beta_i|\phi_{l'}\rangle = \langle\phi_l(\alpha_i)|\alpha_i\beta_i|\phi_{l'}(\beta_i)\rangle = \int\int \alpha_i\beta_i \phi_l^*(\alpha_i)\phi_{l'}(\beta_i)d\alpha_i d\beta_i \tag{27}$$

$\alpha$ and $\beta$ are Cartesian components labled $(x, y, z)^T$ with atom A being centered in $\boldsymbol{R}_A = (x_i, y_i, z_i)^T$ where i is a form of pointer/label dereferencing. $\delta_{\alpha\beta}$ is just the delta function, i.e is is 1 if $\alpha$ and $\beta$ are the same label and 0 otherwise, this serves to include the term only for the diagonal.

$$\boldsymbol{\Theta}_A = \begin{pmatrix} \Theta_A^{xx} & \Theta_A^{xy} & \Theta_A^{xz} \\ \Theta_A^{yx} & \Theta_A^{yy} & \Theta_A^{yz} \\ \Theta_A^{zx} & \Theta_A^{zy} & \Theta_A^{zz} \end{pmatrix} \tag{28}$$

$$\boldsymbol{\mu}_A = \begin{pmatrix} \mu_A^x & \mu_A^y & \mu_A^z \end{pmatrix}^T \tag{29}$$

$$\boldsymbol{R}_{AB} = \boldsymbol{R}_A - \boldsymbol{R}_B \tag{30}$$

$$R_{AB} = \sqrt{(\boldsymbol{R}_{AB}^x)^2 + (\boldsymbol{R}_{AB}^y)^2 + (\boldsymbol{R}_{AB}^z)^2} \tag{31}$$

$$f_n(R_{AB}) = \frac{f_{damp}(a_n, R_{AB})}{R_{AB}^n} = \frac{1}{R_{AB}^n}\frac{1}{1 + 6\left(\frac{R_0^{AB}}{R_{AB}}\right)^{a_n}} \tag{32}$$

$$R_0^{AB} = 0.5(R_0^{A\prime} + R_0^{B\prime}) \tag{33}$$

$$R_0^{A\prime} = \begin{cases} R_0^A + \frac{R_{max} - R_0^A}{1 + exp[-4(CN_A' - N_{val} - \Delta_{val})]} & \text{if } N_{val} \text{ is given} \\ 5.0 \text{ bohrs} & \text{otherwise} \end{cases} \tag{34}$$

$$R_{max} = 5.0 \text{ bohrs} \tag{35}$$

$$\Delta_{val} = 1.2 \tag{36}$$

$R_0^A$ is a fitted value for 12 elements and 5.0 for the rest. $a_n$ are adjusted global parameters. Where $f_{XC}^{\mu_A}$ and $f_{XC}^{\Theta_A}$ are fitted values.

### 3.3 Dispersion contribution

$$F_{\mu\nu}^{D4} = -\frac{1}{2} S_{\mu\nu}(d_A + d_B), \forall \mu \in A, \nu \in B \tag{37}$$

$$\begin{aligned} d_A = & \sum_{r}^{N_{A,ref}} \frac{\partial \xi_A^r(q_A, q_{A,r})}{\partial q_A} \sum_{B} \sum_{s}^{N_{B,ref}} \sum_{n=6,8} \\ & W_A^r(CN_{cov}^A, CN_{cov}^{A,r}) W_B^s(CN_{cov}^B, CN_{cov}^{B,s}) \xi_B^s(q_B, q_{B,s}) \times \\ & s_n \frac{C_n^{AB,ref}}{R_{AB}^n} f_n^{damp,BJ}(R_{AB}) \end{aligned} \tag{38}$$

The dispersion coefficient for two reference atoms $C_n^{AB,\mathrm{ref}}$ is evaluated at the reference points, i.e., for $q_A = q_r$, $q_B = q_s$, $CN_{\mathrm{cov}}^A = CN_{\mathrm{cov}}^r$, and $CN_{\mathrm{cov}}^B = CN_{\mathrm{cov}}^s$.

The Gaussian weighting for each reference system is given by:

$$W_A^r(CN_{cov}^A, CN_{cov}^{A,r}) = \sum_{j=1}^{N_{gauss}} \frac{1}{\mathcal{N}} \exp\left[-6j \cdot (CN_{cov}^A - CN_{cov}^{A,r})^2\right] \tag{39}$$

with

$$\sum_{r}^{N_{A,ref}} W_A^r(CN_{cov}^A, CN_{cov}^{A,r}) = 1 \tag{40}$$

$\mathcal{N}$ is a normalization constant.

$$\mathcal{N} = \sum_{A,\mathrm{ref}=1}^{N^{A,\mathrm{ref}}} \exp\left[-6j \cdot (CN^A - CN^{A,\mathrm{ref}})^2\right] \tag{41}$$

// Write r or ref? CN with or without cov?

The number of Gaussian function per reference system $N_{gauss}$ is mostly one, but equal to three for $CN_{cov}^{A,r} = 0$ and reference systems with similar coordination number.

$C_6^{AB}$ is the pairwise dipole-dipole dispersion coefficients calculated by numerical integration via the Casimir-Polder relation.

$$C_6^{AB} = \frac{3}{\pi} \sum_j w_j \overline{\alpha}_A(i\omega_j, q_A, CN_{cov}^A) \overline{\alpha}_B(i\omega_j, q_B, CN_{cov}^B) \tag{42}$$

$w_j$ are the integration weights, which are derived from a trapeziodal partitioning between the grid points $j \in \{2, \ldots, 22\}$.
The isotropically averaged, dynamic dipole-dipole polarizabilites $\overline{\alpha}$ at the $j$th imaginary frequency $i\omega_j$ are obtained from the self-consistent D4 model; i.e.,

they are depending on the covalent coordination number and are also charge dependent.

$$\overline{\alpha}_A(i\omega_j, q_A, CN^A_{cov}) = \sum_r^{N_{A,ref}} \xi^r_A(q_A, q_{A,r})\overline{\alpha}_{A,r}(i\omega_j, q_{A,r}, CN^{A,r}_{cov})W^r_A(CN^A_{cov}, CN^{A,r}_{cov}) \tag{43}$$

$$\overline{\alpha}_{A,r}(i\omega_j, q_{A,r}, CN^{A,r}_{cov}) = \sum_{A,\mathrm{ref}=1}^{N^{A,\mathrm{ref}}} \alpha^{A,\mathrm{ref}}(i\omega, q_A)W^r_A \tag{44}$$

The charge-dependent atomic dynamic polarizability for a single reference system of atom A is given by the product of $\alpha^{A,\mathrm{ref}}(i\omega)$ and its scaling function as:

$$\alpha^{A,\mathrm{ref}}(i\omega, q_A) = \alpha^{A,\mathrm{ref}}(i\omega)\xi^r_A(q_A, q_{A,r}) \tag{45}$$

$$\alpha^{A,\mathrm{ref}}(i\omega) = \frac{1}{m}\left[\alpha^{AmXn}(i\omega) - \frac{n}{l}\alpha^{X_l}(i\omega)\xi^r_A(q_X, q_{X,r})\right] \tag{46}$$

// The effective nuclear charges $z^{X,\mathrm{ref}}$ entering equation 46 are constant values determined once for the respective reference system. (Find out how to get them)

The charge-dependency is included via the empirical scaling function $\xi^r_A$.

$$\xi^r_A(q_A, q_{A,r}) = \exp\left[3\left\{1 - \exp\left[4\eta_A\left(1 - \frac{Z^{eff}_A + q_{A,r}}{Z^{eff}_A + q_A}\right)\right]\right\}\right] \tag{47}$$

where $\eta_A$ is the chemical hardness taken from ref 98.
$Z^{eff}_A$ is the effective nuclear charge of atom A.

$C^{AB}_8$ is calculated recursively from the lowest order $C^{AB}_6$ coefficients.

$$C^{AB}_8 = 3C^{AB}_6\sqrt{\mathcal{Q}^A\mathcal{Q}^B} \tag{48}$$

$$\mathcal{Q}^A = s_{42}\sqrt{Z^A}\frac{\langle r^4\rangle^A}{\langle r^2\rangle^A} \tag{49}$$

$\sqrt{Z^A}$ is the ad hoc nuclear charge dependent factor.

From the original xTB program we can see that $s_{42}$ is 0.5, and $Z^A$ is the atomic number of A.

$$\sqrt{0.5\left(\frac{r^4}{r^2}\sqrt{Z^A}\right)} \tag{50}$$

$\langle r^4 \rangle$ and $\langle r^2 \rangle$ are simple multipole-type expectation values derived from atomic densities which are averaged geometrically to get the pair coefficients.

$CN^A_{cov}$ is the covalent coordination number for atom A.

$q$ is the atomic charge, so $q_A$ is the atomic charge for atom A.

The scaling parameters in the dispersion model are:

$$a1 = 0.52 \quad | \quad a2 = 5.0 \quad | \quad s6 = 1.0 \quad | \quad s8 = 2.7$$

BJ = Becke-Johnson

$$f_n^{damp,BJ}(R_{AB}) = \frac{R_{AB}^n}{R_{AB}^n + (a_1 \times R_{AB}^{crit} + a_2)^6} \tag{51}$$

$$R_{AB}^{crit} = \sqrt{\frac{C_8^{AB}}{C_6^{AB}}} \tag{52}$$

$$f_9^{damp,zero}(R_{AB}, R_{AC}, R_{BC}) = \left( 1 + 6 \left( \sqrt{\frac{R_{AB}^{crit} R_{BC}^{crit} R_{CA}^{crit}}{R_{AB} R_{BC} R_{CA}}} \right)^{16} \right)^{-1} \tag{53}$$

# 4 Total Energy for GFN2-xTB

$$\begin{aligned} E_{GFN2-xTB} &= E_{rep}^{(0)} + E_{disp}^{(0,1,2)} + E_{EHT}^{(1)} + E_{IES+IXC}^{(2)} + E_{AES+AXC}^{(2)} + E_{IES+IXC}^{(3)} \\ &= E_{rep} + E_{disp}^{D4'} + E_{EHT} + E_\gamma + E_{AES} + E_{AXC} + E_\Gamma^{GFN2} \end{aligned} \tag{54}$$

## 4.1 Repulsion Energy

$$E_{rep} = \frac{1}{2} \sum_{A,B} \frac{Z_A^{eff} Z_B^{eff}}{R_{AB}} e^{-\sqrt{a_A a_B}(R_{AB})^{(k_f)}} \tag{55}$$

$$k_f = \begin{cases} 1 & if A, B \in \{\mathrm{H}, \mathrm{He}\} \\ \frac{3}{2} & otherwise \end{cases} \tag{56}$$

$Z^{eff}$ and $a$ are variables fitted for each element. A,B are the labels of atoms. Since we only have C and H in our systems we can simplify this quite a bit in code. $R_{AB}$ is the distance between the A and B atoms.

## 4.2 Extended Hückel Theory Energy

$$E_{EHT} = \sum_{\mu\nu} P_{\mu\nu} H^{EHT}_{\mu\nu} \tag{57}$$

$$P_{\mu\nu} = P^0_{\mu\nu} + \delta P_{\mu\nu} \tag{58}$$

$$P^0 = \sum_A P^0_A \tag{59}$$

$$\delta P_{\mu\nu} = ?? \quad \text{comes from the iteration, can be skipped for now} \tag{60}$$

Where $P^0_A$ is the neutral atomic reference density of A. This is known as Superposition of Atomic Densities or SAD.

## 4.3 Isotropic electrostatic and Exchange-correlation energy

### 4.3.1 Second order

$$E_\gamma = \frac{1}{2} \sum_{A,B}^{N_{atoms}} \sum_{l\in A} \sum_{l'\in B} q_{A,l} q_{B,l'} \gamma_{AB,ll'} \tag{61}$$

### 4.3.2 Third order

$$E^{GFN2}_\Gamma = \frac{1}{3} \sum_{A}^{N_{atoms}} \sum_{l\in A} (q_{A,l})^3 \Gamma_{A,l} \tag{62}$$

## 4.4 Anisotropic electrostatic energy

$$\begin{aligned} E_{AES} &= E_{q\mu} + E_{q\Theta} + E_{\mu\mu} \\ &= \frac{1}{2} \sum_{A,B} \{ f_3(R_{AB}) [q_A(\boldsymbol{\mu}_B^T \boldsymbol{R}_{BA}) + q_B(\boldsymbol{\mu}_A^T \boldsymbol{R}_{AB})] \\ &\quad + f_5(R_{AB}) [q_A \boldsymbol{R}_{AB}^T \boldsymbol{\Theta}_B \boldsymbol{R}_{AB} + q_B \boldsymbol{R}_{AB}^T \boldsymbol{\Theta}_A \boldsymbol{R}_{AB} \\ &\quad - 3(\boldsymbol{\mu}_A^T \boldsymbol{R}_{AB})(\boldsymbol{\mu}_B^T \boldsymbol{R}_{AB}) + (\boldsymbol{\mu}_A^T \boldsymbol{\mu}_B) R^2_{AB}] \} \end{aligned} \tag{63}$$

## 4.5 Anisotropic XC energy

$$E_{AXC} = \sum_A (f^{\mu_A}_{XC} |\boldsymbol{\mu}_A|^2 + f^{\Theta_A}_{XC} ||\boldsymbol{\Theta}_A||^2) \tag{64}$$

What norms are these?

## 4.6 Dispersion Energy

$$
\begin{aligned}
E_{disp}^{D4'} = &- \sum_{A>B} \sum_{n=6,8} s_n \frac{C_n^{AB}(q_A, CN_{cov}^A, q_B, CN_{cov}^B)}{R_{AB}^n} f_{damp,BJ}^{(n)}(R_{AB}) \\
&- s_9 \sum_{A>B>C} \frac{(3cos(\theta_{ABC})cos(\theta_{BCA})cos(\theta_{CAB}) + 1) C_9^{ABC}(CN_{cov}^A, CN_{cov}^B, CN_{cov}^C)}{(R_{AB}R_{AC}R_{BC})^3} \\
&\times f_{damp,zero}^{(9)}(R_{AB}, R_{AC}, R_{BC}).
\end{aligned} \tag{65}
$$

The term in the second line is the three-body Axilrod– Teller–Muto (ATM) (What is this??????) term and the last line is the corresponding zero-damping function for this term.

The damping and scaling parameters in the dispersion model are:

$$s6 = 1.0 \quad | \quad s8 = 2.7 \quad | \quad s9 = 5.0$$

$C_9^{ABC}$ is the triple-dipole constant[1]:

$$C_9^{ABC} = \frac{3}{\pi} \int_0^\infty \alpha^A(i\omega)\alpha^B(i\omega)\alpha^C(i\omega)d\omega \tag{66}$$

The three-body contribution is typically $< 5-10\%$ of $E_{disp}$, so it is small enough that we can reasonably approximate the coefficients by a geometric mean as[1]:

$$C_9^{ABC} \approx -\sqrt{C_6^{AB}C_6^{AC}C_6^{BC}} \tag{67}$$

$\theta_{ABC}$ is the angle between the two edges going from B to the other two atoms. $\theta_{BCA}$ is the angle between the edges going from C to the other two and so on.

## 4.7 SAD - Superposition of Atomic Densities

The superposition of atomic densities(SAD) is an approach to obtain a good approximation of a collection of atoms, to be used as an initial guess for solving the self-consistent field(SCF) equation.

As originally implemented in DISCO, the molecular electron density can be obtained by adding the densities of all the constituting atoms.

This is how we get the density matrix for an isolated atom? equation 15 from: (https://sci-hub.box/10.1002/jcc.540030314)

$$D_{ij} = \sum_a^{occ} c_{ia}c_{ja} \tag{68}$$

[1] https://www.researchgate.net/publication/43347348_A_Consistent_and_Accurate_Ab_Initio_Parametrization_of_Density_F
D_for_the_94_Elements_H-Pu

To get the coefficients we need to solve SCF for each atom? this is supposedly cheap, but idk how to do it. (https://sci-hub.box/10.1002/jcc.20393) Though the math for Direct SCF Approach is given in this paper at equation 10: (https://sci-hub.box/10.1002/jcc.540030314). This is probably how.

The SAD method is then the sum of all of these?

Equation 2 in the GFN2 paper talks about "superposition of (neutral) atomic reference densities". Is this relevant?

Direct SCF Approach

$$\begin{aligned}\Delta F_{ab} =&(c_{ia}c_{jb} + c_{ja}c_{ib})\\ &\Delta F_{ij} + (c_{ia}c_{kb} + c_{ka}c_{ib})\\ &\Delta F_{ik} + (c_{ia}c_{lb} + c_{la}c_{ib})\\ &\Delta F_{il} + (c_{ja}c_{kb} + c_{ka}c_{jb})\\ &\Delta F_{jk} + (c_{ja}c_{lb} + c_{la}c_{jb})\\ &\Delta F_{jl} + (c_{ka}c_{lb} + c_{la}c_{kb})\Delta F_{kl}\\ &= l_{ijkl}(4E^{ab}_{ij}D_{kl} + 4D_{ij}E^{ab}_{kl} - E^{ab}_{ik}D_{jl} - D_{ik}E^{ab}_{jl} - E^{ab}_{il}D_{jk} - D_{il}E^{ab}_{jk})\end{aligned} \tag{69}$$

where

$$E^{ab}_{ij} = c_{ia}c_{jb} + c_{ja}c_{ib} \tag{70}$$

Equation 18 from (https://sci-hub.box/https://doi.org/10.1021/acs.chemrev.5b00584) uses $\rho_0$ which is the superposition of neutral atom densities:

$$\rho_0 = \sum_A \rho_0^A \tag{71}$$

# 5 Methodology

## 5.1 Testing

The xTB algorithm computes a lot different energies, corrections and other additions to the energy terms. There is a lot of overlap between the different variants of xTB, so even focusing on just one of them still requires a great amount of code. Dealing with large computations with so many small parts makes proper validation especially important, as it becomes increasingly easy to make mistakes. In the case of the xTB algorithm, the original Fortran implementation becomes an important point of reference for comparison. Throughout the coding process it became apparent that the xTB implementation by the people at Grimme Laboratories does not match the equations described in the original xTB paper by the same authors. It should be noted that the program is described as a semiempirical implementation, which could explain some degree of deviation in results from the original paper. With this realization the obvious approach forward was to follow the semiempirical implementation rather than the purely theoretical version. This choice would allow continued validation against the existing code as a reference.

One of the hurdles from testing against an existing program becomes the lack of transparency regarding the logic that finds place between the initial input and the output presented to the user. Thankfully, the source code is publicly available allowing for easy manipulation of the original flow of execution, thus avoiding the hassle of testing against a black box or the need to resort to methods of reverse engineering.

One behalf of these considerations it was decided to write patches that allow to intercept the arguments and results of arbitrary functions just by running the program as normal. This meant that smaller parts could be implemented without the need to implement all the code needed to compute its arguments.

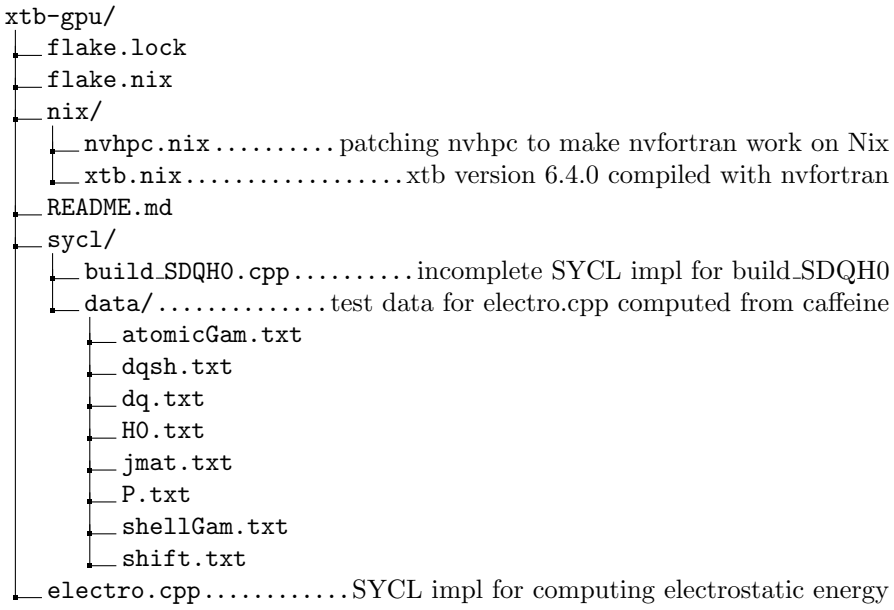
An important part of any software is reproducability, and applying certain patches in certain scenarios is something that should preferably be automated, reproducable, and optimally also portable. This is especially important for this approach to validation as it requires a way to reproduce a specific version of xTB linked to the same versions of dependencies. Essentially an exact copy of the original shell environment to ensure that patches work, results are the same, and no new bugs appear in the program itself or its dependencies. All of this should be achievable without having to add, remove, downgrade or upgrade system packages on your system.

The well-known contenders for this is any of the numerous containerization solutions on Linux, such as Docker, Podman, LXC etc. There are some problems with these options though, one being that it can be difficult to truly reproduce package versions without saving the resulting container image, another being that it does not solve the problem of having multiple versions of the same package installed. Some other notable limitations are that it limits the process to run within the container and passing in a GPU or other hardware can be nefariously difficult. Lastly, a container does not have access to the X or Wayland session needed to run GUI applications, though that is not currently relevant in this case.

Another approach which has been growing in popularity in recent years are specialized package managers that are designed to make reproducable packages, shell environments, systems and other forms of "outputs". Two such popular package managers are Nix from the Nix team and Guix from the GNU foundation. Nix is arguably the more popular option and it is also the solution that has been chosen for this project.

# 6 Code Structure

```
xTB-math/
├── bin2xyz/
│   ├── bin2xyz.cpp ..... converter from float64 coords into xyz format
│   └── C200_10000_fullerenes.float64 .....3d coords for 10k fullerenes
├── flake.lock
├── flake.nix .....conventional structure for Nix inputs and outputs
├── nix/ ..... Nix package definitions for xTB and its dependencies
│   ├── cpx.nix ..... CPCM-X - a solvation model
│   ├── numsa.nix .....solvent accessible surface area calculation
│   ├── patches/ ..... patches for extracting data for validation
│   │   ├── dftd4/ .....dispersion correction
│   │   │   ├── log_args_and_outputs.patch
│   │   │   └── use_gfn2.patch
│   │   └── xtb/
│   │       ├── log_args_and_outputs.patch
│   │       ├── log_electro.patch
│   │       └── log_utils.patch
│   └── xtb.nix ..... extended tight-binding program
├── README.md
├── report/
└── xtb-python/ ..... Python port of xTB paper and Fortran impl
    ├── basisset.py
    ├── blas.py
    ├── cmp_impls.py ..... validation tests against Fortran impl
    ├── data/
    │   ├── C200.xyz
    │   └── caffeine.xyz
    ├── dftd4.py ..... computation for dispersion correction
    ├── dftd4_reference.py ..... constants for dftd4
    ├── energy.py .....various energy computations
    ├── fock.py ..... Fock matrix computation
    ├── gfn2.py ..... xTB-GFN2 specific constants
    ├── lapack.py ..... Facade for LAPACK functions
    ├── scc.py .....computation for self-consistent charges
    ├── slater.py .....computation for slater determinants
    ├── util.py
    └── xyz_reader.py
```


xtb-gpu/
flake.lock
flake.nix
nix/
nvhpc.nix ..... patching nvhpc to make nvfortran work on Nix
xtb.nix ..... xtb version 6.4.0 compiled with nvfortran
README.md
sycl/
build_SDQH0.cpp ..... incomplete SYCL impl for build_SDQH0
data/ ..... test data for electro.cpp computed from caffeine
atomicGam.txt
dqsh.txt
dq.txt
H0.txt
jmat.txt
P.txt
shellGam.txt
shift.txt
electro.cpp ..... SYCL impl for computing electrostatic energy